FOCUSED INFORMATION CRITERIA FOR SELECTING AMONG PARAMETRIC AND NONPARAMETRIC MODELS

by

MARTIN JULLUM

THESIS

presented for the degree of

MASTER OF SCIENCE

(Modelling and Data Analysis)

Statistics Division, Department of Mathematics
Faculty of Mathematics and Natural Sciences
University of Oslo

September 2012

Preface

The long journey ending in this master thesis started out when I bought the book Claeskens and Hjort (2008) on model selection in the summer of 2010. Reading the first chapter introducing model selection with a few practical examples, definitely gave me the impression that model selection was a more important theme of statistics than I was aware of at that time. By browsing the succeeding chapter it was also made clear that still being motivated by purely applied statistics, there was hardcore mathematical statistics underneath – precisely the combination I was looking for to my master thesis. I contacted prof. Hjort, and through a meeting where several possible themes for a master thesis were presented, the theme of nonparametrics vs. parametrics definitely caught my attention. Questions like: "Why aren't there a criterion for selecting between nonparametrics and parametrics?" and "Why haven't anyone thought of the approach via mse estimation before?" were popping into my head, and it was for sure with great endorsement I decided to go for such a theme.

It was demanding to start working with the thesis as I had limited knowledge about both model selection and asymptotic theory. With me officially starting on the thesis in January 2011 and the university's versions of these courses being taught in the spring and autumn that year, I had to study most of the material by myself to get started. That being said, both these courses where excellent, and especially the course on asymptotic theory gave me the basic background necessary to be able to fully understand the more technical mathematical statistics which I needed for the thesis. Even though I started with rather blank sheets, I was rather quickly able to reach an asymptotic result providing estimators for the mean squared error and thus a FIC scheme for a few special cases. The most difficult work of the thesis has consequently been to generalize those results, to obtain nice conditions and to expand them to similar situations. For this work the excellent, but rather theoretical books of Shao (2003) and van der Vaart (2000) has been very helpful.

The situation of iid data has been the main focus of the thesis all along. Clearly most time and effort have been spent handling this situation. This has made chapter 3 which handles this type of data the definite main chapter with the most profound investigations and results. Since regular iid data have been of main interest, chapter 4 on the censored data analogue is not as fine-tuned and detailed regarding sufficient conditions and its implications as the preceding one. This is just a result of the choice that investigations, beyond deriving and stating a fully working general FIC scheme, have been given less priority than other themes more closely connected to the main iid situation. For instance chapter 5 is meant as a collection of related themes with a varying degree of completeness. This priority choice has however caused this thesis to be unusually long for a master thesis, since we are touching so many topics related to the main theme. Unfortunately, there were simply not enough time to follow all of the ideas for these related topics to the very end. On the other hand, the focus on iid data and the

situation in chapter 3 has resulted in the main criterion being studied deeply, while at the same time the useful related topics were not forgotten. Although a thesis like this should be rather theoretical in spirit, my intention has been that since the fundament is motivated by applications, the applied side should neither be ignored. Therefore a great amount of time has been spent developing and building a user friendly function in `R` which makes it possible to apply FIC schemes for quite general situations of iid data. Through the thesis this function is applied to real data examples in order to give an applied touch on the thesis as well.

Originally this thesis was supposed to be finished a few months earlier. In the late spring of 2011 I decided to postpone the deadline for delivering my master thesis from the end of spring 2012 to the end of autumn 2012 and study at 2/3 of full time the last year. This was done to be able to prepare and hopefully to do well in the European and World Championships in Trail Orienteering in May and June 2012, and at the same time be able to complete a thesis that I could be 100 % satisfied with. However, as spring came the following year, a really interesting PhD position in statistics was advertised. To get the position it was required that I should finish my degree some months earlier than I had postponed it to – so I therefore expedited the deadline again. This caused the last couple of months as a master student to be very intense in order to finish the thesis in time.

For me it has been a major goal to do something appreciated with my master thesis, something theoretical oriented possibly with a practical background. Looking back on the product of this thesis I am truly delighted with the obtained results. For this I am sincerely grateful to my supervisor prof. Nils Lid Hjort for introducing me to this very interesting theme. Even though supervising over e-mail the full year when prof. Hjort had a sabbatical year at the University of Cambridge was quite challenging for both of us, I am deeply indebted for his guidance and helpful discussions. This somewhat uncomfortable situation also resulted in me being forced to work more independently, which probably have caused me to learn even more.

In addition I would like to thank my fellow students at reading room *b800* for making the everyday at the university a pleasure. Finally I would like to thank my family, roommates and other friends for their support, and especially Marit with her mentally encouragement and love from almost 400 kilometers away.

Oslo, September 2012,
Martin Jullum

Contents

Chapter 1

Introduction

This chapter contains a short background for the theme of the thesis, as well as an outline of what we attempt to achieve by our thesis. We keep this part rather untechnical to ease the first meeting with theme. In the end of the chapter we give some notes on the notation.

1.1 Background

Given any finite data set where it is natural to assume that the data originate from a common but unknown distribution, the statistician's natural approach to investigate and conclude from these data, is to assume some (possibly approximate) known model for the data. This strategy has turned out to work pretty well for hundreds of years and important decisions have been based upon such investigations. The range of accessible models to fit has however increased dramatically over the years and there are really no limitations on the number of different models it is reasonable to fit. So which of these models should we use? Model selection is, as the name reveals, the step of the statistical analysis where the model(s) for further investigation and conclusions are selected.

Model selection was not much of a field in statistics just a generation ago. This is mainly due to the fact that it was a comprehensive task just to fit *one* model to a data set a few decades ago, and one then often settled with the model one was able to fit. Nowadays one can however fit lots of models in a few seconds with any computer. Model selection has broadened to become a common part of a statistician's task after it gained great acknowledgment after the famous invention of Akaike in 1971. Akaike developed, and published a few years later (Akaike (1974)) an information criterion (AIC) that could be used to select among a number of parametric models. The criterion was originally developed for time series models , but was early on also applied to any other likelihood model. Following the success of this popular criterion, a new field of statistics appeared and numerous alternatives were developed. Among the most famous are the Bayesian approach BIC (Schwarz (1978)), the bias corrected version for linear regression and autoregressive models AICc (Sugiura (1978)) and the model robust version TIC (Takeuchi (1976)). Empirical techniques like cross-validation introduced by Stone (1974) and Geisser (1975) have also been used extensively in some fields, for model validation and selection.

All of the above mentioned criteria are inference independent and chooses model exclusively based on data. A few years ago a new and somewhat different approach drew attention. A model selection criterion where the objective and goal of succeeding inference was directly

included in the model selection step, were developed. The focused information criterion (FIC) due to Claeskens and Hjort (2003), considers a parameter of interest and attempts to select the model preforming best at estimating this particular parameter. This is performed over a set of parametric models where all models are special cases of the model with the most parameters. The criterion attempts to estimate the mean squared error of the focus parameter under each of the competing models, and the scheme selects the model with the smallest estimated error.

1.2 Theme and structure

The theme of this thesis is to transform the idea from the original FIC over to situations where a nonparametric model is included in the set of competing models in addition to a number of parametric models. Being able to compare nonparametric and parametric models is a property that few other model selection criteria possess. The reason for this is that most information criteria, included those mentioned above, relies on the likelihood of the parametric distribution, and most nonparametric models do not have any likelihood, at least not in the same sense as parametric models. Goodness of fit testing based on nonparametrics may in some sense be seen as model selection even if that is not the intention of the test. Disregarding this approach, there are no fully working selecting schemes which compare parametrics and nonparametrics directly, as far as we are aware of. At least it was an unexplored idea to approach the comparison of parametrics and nonparametrics from the focused model selecting perspective, as the work on this thesis began. Nevertheless, it should be noted that the unpublished report Tarima (2011) has some thoughts similar to our approach.

The governing idea that the main part of this thesis is built on, is the same as for the original FIC. In other words, most criteria in this thesis are based on an attempt to minimize the mean squared error (mse) for estimators of a focus parameter. The FIC routines of this thesis select the model that has the smallest estimated mse. We use different techniques to estimate the mse, but most techniques are based on large sample properties of the model estimators.

The thesis is outlined in a somewhat unusual way by reaching the peak and main result rather early on, while the rest of the thesis is spent on investigating implications of the main result and treating similar situations. The thesis start with the most important and comprehensive chapter where FIC in the iid setting is investigated. FIC schemes in other interesting settings along with related topics are treated in later chapters. Some of them are carried out to the full extent, while others are given less time and effort. The appendices are also rather comprehensive to avoid filling up the space in the main thesis and keep the reader focused on what is new theory and what that is just restated results.

In the following and consequently in the whole thesis, we speak about Focused Information Criteria (FIC) to mean the information criteria of this chapter, and not the criterion developed in Claeskens and Hjort (2003). When referring to the criterion of Claeskens and Hjort, such will be emphasized and sometimes denoted "the original FIC". We also stress already here that even if criteria and other results are presented with only one formula for parametric models and estimators, everything is applicable with several parametric models as long as all assumptions holds for each of them and nothing else is stated. This is done completely for ease of presentation.

Even if the reader is not interested in derivations, some knowledge about basic statistics must be held to follow the basic arguments of the thesis. Knowledge about themes as randomness, expectation, variance, covariance and independence, in addition to knowing what a

hypothesis test and a parametric model are, will be assumed and not dealt with here. In addition, a mathematically tuned mind is preferred to fully understand what is going on. With that being said the principles should be accessible for a wide audience. The rest of this introduction contains a chapter by chapter overview of the content of the thesis.

Since this thesis uses bits of pieces of theory from quite many fields in mathematical statistics, chapter 2 is granted to an introduction to the required topics from the fields of statistics central for the thesis. In addition this chapter contains a more detailed review of the most important model selection routines available. Note however that themes necessary just for one single chapter will be introduced when needed in the thesis.

Chapter 3 is the main chapter of thesis. The chapter concerns FIC in the most common situation in statistics, where data are assumed to be independent and identically distributed (iid) scalar variables. The chapter starts out by presenting a few assumptions and based on these we derive the master lemma – a lemma containing the joint limiting distributions of the estimators of the focus parameter. We then use the lemma to obtain approximate estimates for the mse and define these as FIC scores. Thereafter we state and discuss sufficient conditions for the underlying assumptions to hold and prove consistency for the estimators included in the FIC formulae. Furthermore, we investigate how the scheme tends to select models as the sample size increases under different assumptions of the true distribution. Moreover, we slightly touch the art of comparing FIC with other information criterion in terms of performance, and discuss a certain untraditional use of the derived scheme. Finally, we present a multivariate extension of the apparatus and finish off by giving a few examples and illustrations based on data.

In chapter 4 we treat FIC for iid data which are censored. We start out by introducing theory of stochastic process and survival analysis. We then state some working conditions and a lemma with the joint limiting distribution similar to the one in the preceding chapter. We do however specialize on the two most common focus parameters for censored data: the cumulative hazard function and the survival function. In the same manner as in the previous chapter, we use this lemma to obtain approximate estimates for the mse and define FIC schemes based on these. Sufficient conditions are then discussed in addition to the expansion to more general focus parameters. We finish off by providing simplified formulae for a certain special case.

Chapter 5 is devoted to less detailed treatment of various topics in the world of statistics where focused model selection between nonparametrics and parametric may be of interest. Firstly we discuss FIC for density estimation and for regression, where the nonparametric estimators in both situations are based on kernel functions. We further discuss FIC for focus parameters based on two samples in a general setting. A FIC scheme similar to that of chapter 3 is then presented when working under a local misspecified framework similar to that of the "original FIC". Towards the end we roughly discuss FIC based exclusively on resampling techniques and not plug-in estimators in addition to FIC for a parametric model not fitting with the theory of chapter 3.

Chapter 6 concerns weighted FIC (wFIC). In quite general terms we present a model selection scheme where the focus is not primarily on one single focus parameter, but may depend on several focus parameters simultaneously in terms of some weight function. We then discuss how a certain special case of wFIC is connected to a certain goodness of fit test. We finish off by applying a wFIC scheme to a data example.

The last paths of theoretical ideas are presented in chapter 7. The chapter concerns model averaging where the final estimator one use for further inference is based not only on one model,

but is a weighted average of several estimators under different models. We first introduce the concept and present model averaging schemes based on other selection criteria. We then suggest a model averaging scheme whose weight function are related to the FIC schemes presented in this thesis. Moreover we derive the limiting distribution of the final estimator under a few assumptions. We finally apply model averaging to an example.

Chapter 8 contains a brief overview and explanation of an `R` function specially programmed to calculate the FIC scores and perform model selection in a general iid situation. We also give a few lines of code showing how the program is meant to be used.

In the last chapter we summarize the content of the thesis and attempt to point to the main achievements. In addition we discuss a few topics for further work.

The thesis also has two appendices. Appendix A contains a derivation of the joint limiting distribution for the nonparametric and parametric focus parameter estimators under a locally misspecified framework. Appendix B is meant to act as an encyclopedia for the definitions and theorems we apply in the thesis. The results are rewritten in the notation of the thesis, and some are simplified to not cause confusion by being much more general than we need in this thesis.

Finally we note that apart from the illustrative code in chapter 9, no computer code is included in the thesis. If we were going to include all code the whole thesis would simply have turned out to massive. The `R` function alone consists of over 1500 lines of code. Instead we have gathered both the source code of the `R` function and all code used for the examples on the web page `http://folk.uio.no/martinju/FIC`.

1.3 Some notes on notation

This section will be used to clarify themes where we differ from the most common terminology, and to introduce the most important notation. Terminology that is not mentioned in this section will be defined at first time use in the thesis. However, what is already standard terminology in statistics will not be mentioned tediously. Abbreviations will be given in parentheses.

In our notation we attempt to be precise, but still not overwhelm the reader with superscripts and subscripts. As far as it has been possible the most common notation of statistics is used here as well. In addition we have attempted to give similar notation to similar quantities. When stating general definitions, theorems, lemmas and corollaries, the notation may however be different from this. This is done mainly to emphasize that the results are general and does not only hold for our particular application. Note that since the thesis handles so many different themes, and we strive to use notation that is familiar to the reader, the same notation may be used for different quantities in different chapter. This does however only occur where there is no connection between the quantities, and we feel confusion is highly unlikely. Such incidents are also kept to a minimum, and do not regard key quantities. Note also that in chapter 4, the notation will differ slightly from the rest of the thesis. The reason for this is that the standard notation are so incorporated into the field that it is simply easier to read the chapter if we adopt the same notation.

Notice that we will not differ in notation between scalars and (column) vectors, as this will be clear from the context. Where it is not obvious, we will emphasize the dimension of the quantities. Otherwise we adopt most standard mathematical operations and especially we use $()^t$ to denote the transpose of matrices.

When working with a data set in this thesis, we will most often denote it as $Y_1, \ldots, Y_n$,

where Y_i for $i = 1, \ldots, n$ is the data "point" (or a vector) i of the data set with sample size n. We call the set where the data takes values the sample space and denote it by Ω. The usual assumption will be that these data are independent identically distributed (iid), from a true distribution with a cumulative distribution function (cdf) denoted by $G(y)$. When the distribution is assumed to be continuous, we say that the probability density function (pdf or simply density) of the data is $g(y)$, and when the distribution is discrete, we say that the probability mass function (pmf) is $g(y)$. One might think of situations where some part of the sample space is continuous and some part is discrete, but since we do not distinguish between these two data types by our notation, this will not create any trouble in terms of notation. $\widehat{G}_n$ will denote the empirical cdf of the data.

When working with parametric distributions, we will be denoting the cdf by $F(y;\theta)$ and the density or pmf will be denoted by $f(y;\theta)$. Here θ is a p-dimensional parameter vector of the distribution, which takes a value in the parameter space Θ. The notation F_θ and f_θ will be used when we do not stress the evaluation point of these functions. Especially F_θ will be used as a specified cdf and measure even if it depends on the value θ. We will also work with a true or more generally "least false" parameter θ_0 as the minimizer of the Kullback–Leibler divergence between the class of functions on the form f_θ and g. Moreover, $\widehat{\theta}_n$ will denote the maximum likelihood estimator of θ. More on these topics may be found in the next chapter.

Convergence of different types will be very important in this thesis. The different convergence types that we will use will be denoted by $\xrightarrow{P}$, $\xrightarrow{a.s.}$ and $\xrightarrow{L}$, and correspond to respectively convergence in probability, almost sure convergence and convergence in law. These convergence types will be defined in the following chapter. When using this notation we will not state explicitly that this happens as $n \to \infty$, since it is implicitly understood from the context. We also adopt the little "o" and big "o" notation $(o(\cdot), O(\cdot))$ for convergence rates of nonstochastic quantities and the stochastic "colleagues" $o_p(\cdot)$ and $O_p(\cdot)$ for convergence in probability. In addition $\stackrel{d.}{=}$ and $\stackrel{eq.}{\sim}$ denotes respectively equality in distribution and asymptotic equivalence. See e.g. Lehmann (1998) for precise definitions.

To denote norms in a vector space, we will use quite standard notation. However, both the Euclidean norm for vectors and the Frobenius norm for matrices will be denoted by $\|\cdot\|$. For scalars, we will use the absolute value sign $|\cdot|$. The supremum norm (also called the uniform norm, the infinity norm and the Chebyshev norm) will be denoted by $\|\cdot\|_\infty$ as it is the limit of the L_p-norm $\|\cdot\|_p$. We will denote the differentiable of a function $S(x)$ by $\dot{S}(x)$ when it is clear which variable the derivative is calculated with respect to. When the derivative is calculated with respect to a variable other than the main one, we denote it by the use of ∂, like $\frac{\partial}{\partial\theta} S(y;\theta)\big|_{\theta=\theta^*}$.

The focus parameter of interest will be denoted by μ, and assumed to be one-dimensional. In many contexts μ will be seen as a functional of the space of cdfs (see the next chapter). $\mu(H)$ will then be the focus parameter calculated under the cdf H. For simplicity we will also use the following notation: $\mu_{\text{true}} = \mu(G), \widehat{\mu}_{\text{np}} = \mu(\widehat{G}_n), \widehat{\mu}_{\text{pm}} = \mu(F_{\widehat{\theta}_n}), \mu_{0,\text{pm}} = \mu(F_{\theta_0})$, where "pm" and "np" denotes respectively parametric and nonparametric distribution. In addition we write $\mu_F = \mu(F_\theta)$ for our convenience.

Even if we treat differentiating rather regularly, integration may however be seen as somewhat unconventionally treated in this thesis. We will mostly be working with integration with respect to a cdf, which is a valid probability measure. In addition we will use the Lebesgue measure and the counting measure, where the former gives usual "$\mathrm{d}x$ integration" and the latter reduces the integration to a sum. See e.g. Schilling (2005) for an introduction to measure and

integration theory. We could have used only Lebesgue and counting measure integration, but as integration with respect to a cdf gives such a nice and general representation of expectations and its relatives, it is preferred here. Especially, we will be writing

$$\begin{aligned} E_H[S(X)] &= \int_{z\in\Omega} S(x)\,\mathrm{d}H(x) \\ &= \begin{cases} \int_{x\in\Omega} S(x)h(x)\,\mathrm{d}x, & \text{if } X \text{ has continuous distribution and} \\ & h \text{ is the density of } Y, \\ \sum_{i:x_i\in\Omega} S(x_i)h(x_i), & \text{if } X \text{ has discrete distribution and} \\ & h(x_i) = Pr\{X = x_i\}, \end{cases} \end{aligned} \tag{1.1}$$

for the expectation of $S(X)$ when X is a random variable assumed to follow a distribution with cdf H, and S a vector function. Note that by this notation, $\mathrm{d}H(x) = (\mathrm{d}H(x_1), \ldots, \mathrm{d}H(x_r))^t$ if x is r-dimensional, i.e. integration is done element wise and we write $\mathrm{d}H(x)$ even if H takes only scalars. The representation in equation (1.1) is advantageous since it gives the possibility to emphasize which distribution the expectation is calculated under, a key feature in this thesis. The variance ($\mathrm{Var}_H(S(Y))$) and the covariance ($\mathrm{Cov}_H(S_1(Y), S_2(Y))$) for the functions S_1 and S_2, are defined in a similar manner. Also for the probability of some event $A(X)$ depending on X, we will use such a representation. By thinking of the probability as the expectation of the indicator that the event occurs, we may write $Pr_H\{A\} = \int_{z\in\Omega} \mathbf{1}_{\{A\}}(x)\,\mathrm{d}H(x)$. When it is perfectly clear which distribution the random variable has, the subscript may be omitted. Even if integration with respect to the cdf is the preferred one, some tasks are better handled by integrating with respect to the measure υ. υ represents the Lebesgue measure when the distribution is continuous, and represent the counting measure whenever it is discrete. Using this terminology, we get

$$\int S(x)\,\mathrm{d}H(x) = \int S(x)h(x)\,\mathrm{d}\upsilon(x),$$

for h the density or pmf of the data, where the additional (x) is used to emphasize the integration variable of the function.

Moreover, to make the representation in the thesis easier to read, we will use the same notation for the same general quantities. Unless otherwise stated, we will use the following notation: x for a general vector, z (or sometimes y) for general scalar, X for a general random vector variable, S for a general vector function, T for a general functional, H for a general cdf and Z_0 for a standard normal distributed variable. As noted before, we also use μ for the focus parameter, whether seen as a functional or not. The quantity V with different superscripts and subscripts will also be reserved to variance and covariance terms related to the focus parameter. Finally, note that we use 0 not only as a scalar, but also as a p-dimensional column vector of zeros, where p is the dimension of θ. It will be clear from the context when it denotes a scalar and when it denotes a vector.

Chapter 2

Model selection and basics of main topics

This chapter's main objective is to introduce the main topics required to read and fully understand the arguments involved in the proofs of the key results of this thesis. In addition an overview of the field of model selection is provided. The introductory part contains partly tentative definitions and some heuristic arguments explaining the role of the defined statistics. The chapter introduces theory in the fields of asymptotic theory, statistical functionals, influence functions and maximum likelihood theory. Readers with good knowledge of these topics and those not interested in the derivations, can be content with just browsing quickly through this introduction, although we encourage all readers to fully read the chapter to get familiar with how we use the topics. For a more fundamental introduction to basic statistics, see any introductory statistical textbook like Rice (2007). For a more rigorous treatment of these topics Lehmann (1998) is recommended for beginners, whereas Shao (2003) or van der Vaart (2000) are recommended for precise treatment of the more advanced topics of these fields.

2.1 Main topics

We now turn to the introduction of the main topics underlying this thesis.

Statistical functionals

We will in this thesis work extensively with statistical functionals, though not in in a very advanced way. Since no hardcore functional analysis is necessary in the thesis, the more advanced theory will neither be included in this introduction. We do however remind the reader of what a functional is, and introduce different notions of a functional derivative.

A functional is in general a map from a function space into its underlying scalar or vector field. The sort of statistical functionals we will deal with here are functions mapping a cdf over to the real line. E.g. with X a random variable with cdf H, the functional T where $T(H) = \int x dH(x) = E_H[X]$ can be thought of as a functional taking H as an argument and mapping it over to the expectation with respect to H. Thus, different cdfs H give different output of the functional. Note also that if the cdf depends on a parameter, say θ, the functional with respect to that cdf may be written as a regular function of this parameter, $T(H_\theta) = T_H(\theta)$, where only θ is allowed to vary.

The derivative of a general functional will be central in this thesis. There exist several non-equivalent definitions of differentiability of a general functional. We will be working with three types of differentiability. Those are Gâteaux, Hadamard and Fréchet differentiability, where Fréchet is strongest and implies Hadamard, which again implies Gâteaux. For a functional being both Gâteaux and Hadamard differentiable, the derivative of either type is the same. For a cdf H, the Gâteaux derivative of a functional T, in the fixed direction Δ for $\Delta \in \{c(H-H^*), H, H^*$ are cdfs ,$c \in \mathbb{R}\}$, is defined as

$$L_H(H - H^*) = \lim_{\lambda \to 0} \frac{T(H + \lambda(H - H^*)) - T(H)}{\lambda},$$

whenever the differentiable exists, which is the case when the limit is finite.

Hadamard differentiability restricts this definition by requiring that the limit also exists for varying direction Δ as long as the direction stabilizes as $\lambda \to \infty$. Formally the variation is dealt with in terms of a norm or more generally a metric. For our use of Hadamard differentiability, the supremum norm $\|S(z)\|_\infty = \sup_z |S(z)|$, where S is a function on $\mathbb{R}$, will be used. The precise definition of Hadamard differentiability is given in the appendix (definition B.1.1, ii).

Fréchet differentiability is also equipped with a norm $\|\cdot\|_*$. It requires that the change from $T(H_j)$ to $T(H)$ in some way has the same speed as $\|H_j - H\|_*$ when H_j is a sequence of cdfs such that $\|H_j - H\|_* \to \infty$. The precise definition of Fréchet differentiability is given in the appendix (definition B.1.1, iii).

Asymptotic theory

With data $Y_1, \ldots, Y_n$, asymptotic (or large sample) theory investigates what happens as the sample size n grows to infinity. A wide range of results in different fields of statistics has developed from this important theory. The famous law of large numbers (theorem B.2.1) and the central limit theorem (B.2.4) are now standard asymptotic results which again are the basis for most statistical inference done today. This is the case since most hypothesis testing and confidence intervals statisticians deal with, are based directly on this theory – for non-statisticians unfortunately often without knowing it. Consequently they are sometimes misused and dealt with as precise results also for small samples. Asymptotic theory is also essential in this thesis, including convergence in probability and law of both data of different types of variables. A sequence of random variables X_n converges to a random variable X in probability ($X_n \xrightarrow{P} X$), if for every $\epsilon > 0$,

$$Pr\{|X_n - X| < \epsilon\} \to 1 \quad \text{as } n \to \infty.$$

A stronger kind of convergence is convergence *almost surely*. A sequence of random variables X_n is said to converge almost surely to a random variable X ($X_n \xrightarrow{a.s.} X$) if

$$Pr\left\{\lim_{n\to\infty} |X_n - X|\right\} = 1.$$

In a somewhat similar way, a sequence of random variables X_n with corresponding cdfs H_n converges in law to X with cdf H ($X_n \xrightarrow{L} X$) if

$$H_n(x) \to H(x) \quad \text{as } n \to \infty \quad \text{at all continuity points } x \text{ of } H.$$

We now introduce the two most important results of asymptotic theory, the law of large numbers and the central limit theorem. The law of large numbers (LLN) states that the mean of n independent identically distributed (iid) data $Y_1, \ldots, Y_n$ converges to their common mean ξ in probability (or stronger almost surely), i.e.

$$\frac{1}{n}\sum_{i=1}^{n} Y_i \xrightarrow{P} \xi \quad \text{(weak form)},$$
$$\frac{1}{n}\sum_{i=1}^{n} Y_i \xrightarrow{a.s.} \xi \quad \text{(strong form)},$$

provided $E_G[|Y_i|] < \infty$. The central limit theorem (CLT) states that

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i - \xi\right) \xrightarrow{L} N(0, \sigma^2),$$

if the variance σ^2 of Y_i is finite. Under additional regularity conditions of the Lindeberg–Feller type, similar results hold for data $Y_{n1}, \ldots, Y_{nn}$ where the distribution of the data may also vary with the sample size n.

Nonparametrics

Nonparametric statistics is the field of statistics where the aim is to do inference with as few assumptions as possible. As the name reveals, one does not fit or use parameters in any predefined distribution function, but performs inference without assumptions regarding the form of the distribution. The empirical distribution function (ecdf) of a data set $Y_1, \ldots Y_n$, is maybe the most important function in nonparametric statistics. It is a valid cdf given by

$$\widehat{G}_n(y) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{Y_i \leq y\}}(y),$$

i.e. a monotone step function with jumps of size $\frac{1}{n}$ at every data point, used to estimate the true cdf via nonparametrics.

In the multivariate case where $Y_1, \ldots, Y_n$ are all r-dimensional iid variables from the same distribution, one can define a similar estimator. Letting $Y_i = (Y_{i1}, \ldots Y_{ir})$, the ecdf is more generally defined as

$$\widehat{G}_n(y_1, \ldots, y_r) = \frac{1}{n}\sum_{i}^{n} \mathbf{1}_{\{Y_{1i} \leq y_1 \cap \cdots \cap Y_{ri} \leq y_r\}}(y_1, \ldots, y_r),$$

where $\cap$ denotes the intersection of sets, or logical "and".

The ecdf has many great properties. In this thesis we need the property that the ecdf form a nonparametric estimator for any functional parameter μ which may be written as a functional of any cdf H: $\mu = \mu(H)$. Especially $\widehat{\mu}_{\text{np}} = \mu(\widehat{G}_n)$ is the so called plug-in estimator. This estimator has again nice properties under certain regularity conditions. For a smooth enough μ it can be shown in various ways that $\sqrt{n}(\mu(\widehat{G}_n) - \mu(G))$ converges in law to a certain zero-mean normal distribution.

Influence functions

The influence function is a measure of the impact of a change in the underlying distribution of a statistical functional. The influence function can be seen as a special case of the functional derivative and it exists whenever the functional is Gâteaux differentiable in a certain direction. When the influence function exists for a functional μ at the cdf H, it is the linear map given by

$$\mathrm{IF}_\mu(y; H) = L_H(\delta_y - H),$$

where $\delta_y(x) = \mathbf{1}_{\{x \geq y\}}(x)$ is the cdf of Dirac's delta measure assigning mass 1 to the point y. Equivalently, for a function $s : [0,1] \to \mathbb{R}$ given by $s(\lambda) = \mu(F + \lambda(\delta_y - F))$, the influence function may be written as

$$\mathrm{IF}_\mu(y; F) = \dot{s}(\lambda)|_{\lambda=0}.$$

In addition to the fact that the influence function measures the sensitivity with respect to the distribution, it has the property of leading to certain limiting distributions for its functional. Especially it is the main ingredient in a clever way of finding the limiting distribution of the plug-in estimator introduced above.

Another useful property of the influence function, is that by linearity $\dot{\mu}(\widehat{G}_n; H - \widehat{G}_n) = \int \mathrm{IF}_\mu(y; H) d\widehat{G}_n(y) = \frac{1}{n}\sum_{i=1}^n \mathrm{IF}_\mu(Y_i; H)$ for some cdf H. For data $Y_1, \ldots, Y_n$ we define the empirical influence function as $\mathrm{IF}_\mu(y; \widehat{G}_n)$. The function values $\mathrm{IF}_\mu(Y_i; \widehat{G}_n) = \mathrm{IF}_\mu(Y_i; \widehat{G}_n)$ for $i = 1, \ldots, n$ will be of special interest in later sections.

Maximum likelihood theory

The theory of maximum likelihood is very important in modern statistics mainly because of its simple idea and useful properties. For iid data $Y_1, \ldots, Y_n$ taken as realizations from a distribution with density or pmf $f(y; \theta_{\text{true}})$, the likelihood of the data is defined as

$$L_n(\theta) = f_{\text{joint}}(Y_1, \ldots, Y_n; \theta) = \prod_{i=1}^n f(Y_i; \theta).$$

The maximum likelihood estimator (ML estimator) $\widehat{\theta}_n$ is defined as the value of θ that maximizes $L_n(\theta)$. Since the logarithm is a monotone function, it is equivalently defined as the maximizer of $l_n(\theta) = \log L_n(\theta)$. In many situations this representation simplifies the task of finding the value $\widehat{\theta}_n$. Taking the log of the likelihood also leads to other expressions needed to derive the limiting distribution of the ML estimator. We thus define the ML estimator

$$\widehat{\theta}_n = \operatorname*{argmax}_\theta l_n(\theta) = \operatorname*{argmax}_\theta \sum_{i=1}^n \log(f(Y_i; \theta)),$$

the score function

$$U(y; \theta) = \frac{\partial}{\partial \theta} \log(f(y; \theta)),$$

and the information function

$$I(y;\theta) = \frac{\partial^2}{\partial\theta^t\partial\theta}\log(f(y;\theta)) = \frac{\partial}{\partial\theta}U(y;\theta)^t.$$

Under rather mild regularity conditions the ML estimator is consistent, i.e. $\widehat{\theta}_n \xrightarrow{P} \theta_{\text{true}}$. In most situations the true distribution of the data is not known. Neither is it known if the distribution belongs to a certain parametric family with cdf of the form F_θ, for some θ value. ML estimation may however still be performed for this class of distributions. If the true distribution is not part of this parametric class, θ_{true} does not exist or make sense. Instead we then work with the so-called least false parameter θ_0 defined as the θ value that minimizes the Kullback–Leibler divergence[1] between the true, but unknown density or pmf of the data $g(y)$ and the parametric class of with densities or pmfs on the form $f(y;\theta)$. When such a θ_0 exist, we also have $\widehat{\theta}_n \xrightarrow{P} \theta_0$. If it turns out that the data actually do stem from f_θ, it is easily seen that $\theta_0 = \theta_{\text{true}}$. Introducing θ_0 may therefore be seen as a generalization of the standard textbook case where one assumes that the true distribution has density or pmf in the parametric class of f_θ.

In the thesis, the limiting distribution of the ML estimator is central. Under further regularity conditions which are precisely given in theorem (3.3.3), it can be shown that $\sqrt{n}(\widehat{\theta}_n - \theta_0)$ converges in law to a certain zero-mean normal distribution.

2.2 Model selection

Model selection is an important task for a statistician analyzing a data set. Considering parametric models, more parameters means more model flexibility, but greater uncertainty in the estimation process, whereas less parameters means less model flexibility, but more estimation power. A too simple model may not capture a phenomenon important for the later inference, while a too complex model may indicate a nonexistent phenomenon of the data set or have so much uncertainty that conclusions cannot be trusted to the fullest. Thus, choosing a statistical model can be seen as a trading game between model flexibility and uncertainty.

As mentioned in the introduction there exist many different model selection schemes which are easy to use for the statistical researcher and considered mainstreams for statisticians. The information criterion approach is the most widely used method to select a model for data set. Information criteria are characterized by a formula mapping the model and the data over some real number. Depending on whether a big or small value corresponds to a good fit of a model, the schemes choose the model that the criterion value indicate is the best among the candidates.

The first ever information criterion to be published is as mentioned Akaike's information criterion (AIC). It can be applied to any set of parametric models which specifies a likelihood and is defined as

$$\text{AIC}(M_\theta) = 2l_{n,\max} - 2p,$$

where $l_{n,\max} = l_n(\widehat{\theta}_n)$ denotes the maximum of the log-likelihood of the model M_θ, and $p = \dim(\theta)$ is the dimension of the parameter space (or number of univariate parameters). The criterion selects the model amongst the set of candidates whose AIC score is the largest. The

[1] The Kullback–Leibler divergence is a measure of the divergence (loosely speaking a distance) from one distribution to another. The divergence from h_1 to h_2 is defined as $\int h_1(y)\frac{\log(h_1(y))}{\log(h_2(y))}\,\mathrm{d}\upsilon(y)$ for h_1 and h_2 the densities or pmfs of the two distributions.

first term is the main term of the AIC formula specifying how well the model fits, whereas the second is a penalizing term which penalizes for the complexity of the model. Up until asymptotically negligible terms, the AIC score is proportional to a bias adjusted estimator of the decisive ingredient of the expected Kullback–Leibler divergence between the true (unknown) and fitted model. This is one way to motivate AIC.

The Bayesian information criterion (BIC),[2] is due to Schwarz (1978). The BIC criterion,

$$\mathrm{BIC}(M_\theta) = 2l_{n,\max} - \log(n)p,$$

is just like the AIC on the penalized log-likelihood form, and selects the model with the largest score. BIC has as opposed to AIC a penalizing term depending on the sample size n. For large data sets BIC is therefore penalizing more for model complexity than compared to AIC. Since the correction term of AIC does not depend on the sample size, and the fit of the complex models will improve as the sample size increases, compared to simpler models, more and more complex models will be preferred by AIC as n increases. BIC may therefore be a wiser choice of model selection scheme for large data sets. As the name reveals, BIC has a Bayesian motivation.[3] The BIC score is an approximate formula, based on a Laplace integral approximation of the decisive ingredient in the formula for the Bayesian posterior model selection probability when using a flat prior.[4]

Takeuchi's information criterion (TIC), or exact AIC as it is sometimes called, is similar to AIC not only by its formula, but also by its derivation. The criterion, which is due to Takeuchi (1976), is defined as

$$\mathrm{TIC}(M_\theta) = 2l_{n,\max} - 2\widehat{p}^*.$$

Here $\widehat{p}^*$ is an estimator of $p^* = \mathrm{tr}(\mathrm{J}^{-1}\mathrm{K})$, the generalized dimension of the parameter space and $\mathrm{tr}(\cdot)$ denotes the trace of the matrix, i.e. the sum of the diagonal elements of the matrix. The estimator $\widehat{p}^*$ is produced by simply inserting the empirical analogues of J and K. What distinguishes the formulae of AIC and TIC is just the estimator of this p or p^* quantity. In the AIC motivation shown above, the p^* quantity also appears, but the AIC strategy is then to assume for this estimation that the candidate model is the true model, giving $J = K$ and $p^* = p$. TIC, on the other hand, does not rely on this rather unpleasant assumption and uses the data to estimate this quantity. When the candidate model is far from correct, TIC will tend to penalize more than AIC. Note that for high dimensional models there are a lot of extra variables ($p(p+1)$) to estimate when using TIC as a criterion in contrast to AIC. Many extra parameters to estimate causes estimation uncertainty, so when the dimension is high compared to the sample size, the use of TIC is probably not such a good choice after all. When the dimension of the candidate models are small compared to the sample size, TIC should be preferred over AIC.

The corrected AIC (AICc) was first suggested by Sugiura (1978) for linear regression models and later on justified for time series model and other applications by Hurvich and Tsai (1989).

[2] BIC are in sometimes also called the Schwarz information criterion (SIC)

[3] Bayesian statistics is the big counterpart to the traditional type of frequentistic statistics. The Bayesian way of thinking is characterized by thinking of unknowns as being random and having a probability distribution as opposed to the frequentists who think of the unknown variables as fixed.

[4] The posterior is the probability distribution given the data and is often the Bayesian's conclusion after a statistical analysis and the prior is the knowledge about the unknowns before data is considered. A flat prior corresponds to no presumed knowledge.

The criterion has a penalizing term calibrated to work better for small samples than what AIC has. In general terms the criterion may be written as

$$\text{AICc}(M_\theta) = 2l_{n,\max} - 2p\frac{n}{n-p-1},$$

where p still is the total number of parameters in the model. Since AIC is based on asymptotic theory, the use of the criterion is only approximate for finite sample sizes. Especially for small samples sizes Hurvich and Tsai (1989) show that AIC has a large negative bias. The corrected version of AIC attempts to solve this small sample problem by adjusting the penalizing term and let it depend on the sample size. The derivation of AICc differs from application to application and is mostly of a somewhat different style than the AIC motivation referred to above. In the normal linear regression situation, however, AICc might be derived along the same lines as AIC, i.e. by considering the decisive quantity of the expected Kullback–Leibler divergence. Using normality properties known for these regression models one is able to give exact formulae without the use of asymptotical approximations and hence provide more accurate estimates for the particular decisive quantity.

As mentioned in the introduction, a more recent approach of model selection is to include the objective of the forthcoming inference into the model selection step. The focused information criterion (original FIC) of Claeskens and Hjort (2003) base the model selection on the estimation uncertainty of a predefined focus parameter μ. The criterion attempts to estimate the mean squared error of the focus parameter under each of the competing models. Claeskens and Hjort (2003) base their theory on a locally misspecified framework and works out a new criterion selecting among a set of parametric models in both the iid setting and the more general regression setting. The set of parametric models must be on the form where all models are special cases of the model with the most parameters. The criterion may be represented in many different ways. Among them is

$$\begin{aligned}\text{FIC}(M_S) &= \widehat{\text{mse}}_{\text{lim}}(M_S) - c,\\ &= \widehat{\text{Var}}(M_S) + \widehat{\text{bias}^2}(M_S) - c,\\ &= \widehat{\text{Var}}(M_S) + \widehat{\text{bias}}^2(M_S) - \widehat{\text{Var}}\left(\widehat{\text{bias}}(M_S)\right) - c. \end{aligned} \tag{2.1}$$

Here M_S denotes submodel S of the full model denoted by M_{wide}, $\widehat{\text{mse}}_{\text{lim}}(M_S)$ is an estimator of the limiting mean squared error of $\sqrt{n}(\widehat{\mu}_S - \mu_{\text{true}})$ and c is simply a quantity not depending on S. The FIC scheme then naturally selects the model with the smallest $\text{FIC}(M_S)$ value. As the trained eye see from formula (2.1), the FIC scheme estimates the mse as variance plus squared bias. The estimator of the squared bias consists of an estimate of the (non-squared) bias which is squared before an estimate of the variance of this bias estimator finally is subtracted. This criterion is more troublesome since it requires some more calculation and preparation prior to the model selection step. It is therefore somewhat harder to get a grip on compared to the simpler AIC and BIC. As a consequence it is not freely available in computer software packages that statisticians tend to use. This might be the reason it has not yet become a mainstream analysis tool for the group of so called "hobby statisticians", and has mainly been acknowledged by researchers and experts in the field. However, the approach of including the objective of the statistical analysis into the model selection step is "up-and-coming" with an increasing rate of published papers and talks. Examples of published papers applying of the original FIC are Rohan and Ramanathan (2011) using a variant of the criterion for order selection in time series

models and Lien and Shrestha (2005) applying the criterion to estimate optimal hedge ratio in financial mathematics.

As mentioned in the introduction, Tarima (2011) discuss another FIC related topic in a still unpublished paper. The author's idea is to estimate the mse of a quantity of interest by assuming some estimator is approximately unbiased for the true value of this quantity. Pieces of large sample theory and bootstrapping are then used to estimate the mse, unfortunately without additional correction of the squared bias estimate.

There exist lots of other criteria in addition to those already discussed, especially modification of AIC has been popular. In fact most of the letters of the first part of the alphabet has given name to an information criterion. To the already introduced criteria we add the Copula information criterion (CIC) due to Grønneberg and Hjort (2008), the Deviance information criterion (DIC) due to Spiegelhalter et al. (2002) and the Generalized information criterion (GIC) of Konishi and Kitagawa (1996). In addition to this large set of information criteria, there are alternative ways of selecting among competing models for a data set. Cross-validation is a technique where the data set at hand are split into two parts. One is used for fitting a model (also called training) and then the fitted model is used to predict the other part of the data set for validation of the fitted model. The technique of sequentially leaving only one observation out at a time will be the main interest for model selection purposes. Especially, for the iid situation

$$\mathrm{xv}(M) = \frac{1}{n}\sum_{i=1}^{n} \log f_M(Y_i, \widehat{\theta}_{(-i)}),$$

is again an estimator of the decisive quantity of the Kullback–Leibler divergence between the true model and the candidate model. Here $\widehat{\theta}_{(-i)}$ is the maximum likelihood estimator under this particular model, when the i-th data value (or vector or matrix) is left out of the data set. The schemes for regression and other settings follow the same strategy. The model with the largest $\mathrm{xv}(M)$ score is selected. Cross-validation techniques are robust in the sense that one is able to check how well the model operates without bringing in new data. In this model selection situation we only use the technique to produce an estimate of an interesting quantity, but even so, the scheme inherits this prediction robustness. This type of robustness is especially beneficial for small samples or when very complex models are considered. It can furthermore be shown that using this cross-validation scheme as a model selector is first-order large sample equivalent to the use of TIC.

We now turn to indirect model selection via hypothesis testing. The procedure of testing whether a certain regression coefficient is significant or not, is an important step in regression analysis. This may however be seen as model selection. Hypothesis testing is most often based on a normal or χ^2 approximate distribution stemming from some central limit type of theorem, depending on the application type. The null hypothesis is usually that the coefficient tested does not make any difference. The hypothesis is rejected and the covariate included when the large sample theory indicates that the obtained estimates are not due to chance under the null hypothesis, where the boundary between rejection and "acceptance" depends on some significance level α. The significance level is often rather unnaturally set to some value (0.05 is quite common) without any greater reasoning for why exactly this level was chosen. Seen as a model selection method this is not very accommodating.

The last model selection related theme we will discuss here is that of goodness of fit testing. As for hypothesis testing, goodness of fit testing is not primarily thought of as model selection,

but what is done in practice is clearly related to model selection. Goodness of fit testing tests the hypothesis that the data at hand, which usually are iid, stems from a fixed distribution. Pearson's χ^2 test is probably the simplest form of goodness of fit testing. The test consists of splitting the sample space of the fixed distribution into a number of intervals and a comparison of the observed number of samples in each interval against what would be expected under the null hypothesis. The sum of these scaled differences is then compared to a χ^2-distribution. In terms of model selection one would select the fixed model if the p-value is less than the preset significance level, otherwise one should go for nonparametrics. Another quite common test is the Kolmogorov test which compares the maximum distance between the ecdf and the cdf of the fixed distribution, against some χ^2-distribution. Finally there exist tests of the Cramér–von Mises type where $n \int [\widehat{G}_n(x) - F_0(x)]^2 \, \mathrm{d}W(x)$, are used as test statistic. Here $F_0(x)$ is a fixed cdf and $W(x)$ a nondecreasing weight function. More on goodness of fit tests can be found e.g. in Lehmann (1998, chapter 5.7).

As seen above, there exist techniques differing widely in terms of both theoretical justification and practical computation, which may be used to select between different statistical models. Here we have presented some of the most common and general techniques. Especially there exist versions of many of these techniques specially developed to work for certain applications or data types. Data analysis is an enormous field and it may be somewhat optimistic to think that one is able to create a model selection scheme which works very well in all types of applications. Information criteria are however a strategy which is simple in basic theory and can be applied in a wide range of applications. Focused inference and model selection criteria may in this context be seen as a bridge between the generality of information criteria and the specificity of interest driven inference. For a further introduction to model selection techniques see e.g. Claeskens and Hjort (2008, chapter 2,3 and 6).

Chapter 3

FIC for iid data

In this chapter we will work inside what we will refer to as the standard framework, where univariate iid data $Y_1, \ldots, Y_n$ are assumed to originate from a true distribution with density of pmf g and cdf G. This is one of the simplest forms of data a statistical researcher is encountered with, yet still one of the most common ones.

This chapter, handling FIC for the iid data type, is divided into three main parts. The first part consists of lemmas and corollaries containing precise limits of key quantities related to the estimators of the focus parameter μ. The second part and section takes care of estimation based on these results. These estimators create FIC scores and schemes that may be used for model selection between a nonparametric and several parametric models, when we focus on the parameter μ. The third part deals with the consequences and properties of the obtained schemes. Especially, we explore the properties of the estimators forming the FIC scores, and explore the behavior of the scheme under different assumptions about the truth. In addition to these main parts, we propose a multivariate extension of the scheme and give a few examples and illustrations at the end of the chapter.

3.1 Limiting distributions

The approximations used in this thesis are based on the behavior of different parameter estimators and functions of these in the limit. All limiting distributions we are in need of in this chapter can actually be derived from *one* joint limiting distribution. We will therefore start by presenting and deriving this limiting distribution, and then carry out the necessary transformations to arrive at the limiting distributions we shall be using later on.

Before we state the assumptions that we will be working under, let us define a few quantities that play central roles in this chapter. As introduced in section 2.1, $U(y;\theta)$ and $I(y;\theta)$ are respectively the score and information function, and $\mathrm{IF}_\mu(y;G)$ the influence function of μ at G. Furthermore, let

$$\begin{aligned} J &= E_G[I(Y_i;\theta_0)], \\ K &= \mathrm{Var}_G(U(Y_i;\theta_0)), \\ \nu &= \mathrm{Var}_G(\mathrm{IF}_\mu(Y_i;G)), \\ Q &= \mathrm{Cov}_G(U(Y_i;\theta_0), \mathrm{IF}_\mu(Y_i;G)). \end{aligned}$$

Inserting different cdfs in the functional $\mu(\cdot)$ defines the following related quantities:

- $\mu_{\text{true}} = \mu(G)$: The true value of the focus parameter.
- $\mu_{0,\text{pm}} = \mu(F_{\theta_0})$: The least false focus parameter value in the parametric family.
- $\widehat{\mu}_{\text{pm}} = \mu(F_{\widehat{\theta}_n})$: The parametric μ estimator.
- $\widehat{\mu}_{\text{np}} = \mu(\widehat{G}_n)$: The nonparametric μ estimator.

To ease the presentation, let also

$$\overline{U}_n = \frac{1}{n}\sum_{i=1}^{n} U(Y_i;\theta_0),$$
$$\overline{\text{IF}}_{\mu,n}(H) = \frac{1}{n}\sum_{i=1}^{n} \text{IF}_\mu(Y_i;H),$$

for some cdf H.

Assumption 3.1.1. *Let $Y_1,\ldots,Y_n$ be iid variables from a distribution with cdf G. Let μ be a one-dimensional focus parameter, and θ the p-dimensional parameter vector of the parametric family of distributions with cdf F_θ, and θ_0 the unique least false parameter of this parametric family. For this situation assume*

$$\mu(\widehat{G}_n) = \mu(G) + \overline{\text{IF}}_{\mu,n}(G) + o_p\left(\frac{1}{\sqrt{n}}\right), \tag{3.1}$$
$$E_G[\text{IF}_\mu(Y_i;G)] = 0, \qquad E_G[\text{IF}_\mu(Y_i;G)^2] = \nu < \infty, \tag{3.2}$$

and

$$\widehat{\theta}_n = \theta_0 + J^{-1}\overline{U}_n + o_p\left(\frac{1}{\sqrt{n}}\right), \tag{3.3}$$
$$E_G[U(Y_i;\theta_0)] = 0 \qquad E_G[||U(Y_i;\theta_0)||^2] < \infty. \tag{3.4}$$

Finally assume that

$$\left.\frac{\partial \mu_F}{\partial \theta}\right|_{\theta_0} \neq 0. \tag{3.5}$$

These assumptions are rather mild and will hold in most regular cases. Section 3.3 provides sufficient conditions for the key assumptions above to hold. For now, let us take assumption 3.1.1 as true for a given situation and see what results such a situation produces. Below we jump right into the derivation process of the FIC by deriving the main result of this section – the joint limiting distribution of the nonparametric and parametric estimators of μ_{true}. The following lemma provides this limiting distribution:

Lemma 3.1.2. *When the relations and conditions of assumption 3.1.1 hold, the following limiting distribution appears:*

$$\sqrt{n}\begin{pmatrix}\widehat{\mu}_{\text{np}} - \mu_{\text{true}} \\ \widehat{\mu}_{\text{pm}} - \mu_{0,\text{pm}}\end{pmatrix} \xrightarrow{L} N_2\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}V_{\text{np}} & V_{\text{pm,np}} \\ V_{\text{pm,np}} & V_{\text{pm}}\end{pmatrix}\right), \tag{3.6}$$

where

$$\begin{aligned} V_{\text{np}} &= \nu, \\ V_{\text{pm}} &= \left(\left. \frac{\partial \mu_F}{\partial \theta} \right|_{\theta_0} \right)^t J^{-1} K J^{-1} \left(\left. \frac{\partial \mu_F}{\partial \theta} \right|_{\theta_0} \right), \\ V_{\text{pm,np}} &= \left(\left. \frac{\partial \mu_F}{\partial \theta} \right|_{\theta_0} \right)^t J^{-1} Q. \end{aligned}$$

Proof. Assume first that we have shown the following limiting distribution:

$$\sqrt{n} \begin{pmatrix} \widehat{\mu}_{\text{np}} - \mu_{\text{true}} \\ \widehat{\theta}_n - \theta_0 \end{pmatrix} \xrightarrow{L} N_{p+1}\left(0, \Sigma\right), \tag{3.7}$$

where Σ may be written as a block matrix of the form

$$\Sigma = \begin{pmatrix} \Sigma_{00} & \Sigma_{01} \\ \Sigma_{10} & \Sigma_{11} \end{pmatrix},$$

where

$$\begin{aligned} \Sigma_{00} &= \nu, \\ \Sigma_{11} &= J^{-1} K J^{-1}, \\ \Sigma_{10} &= \Sigma_{01}^t = J^{-1} Q. \end{aligned}$$

Let us now apply the delta method (theorem B.2.8) to this limiting distribution with the following transformation function

$$S_\mu(z, x) = \begin{pmatrix} z \\ \mu_F(x) \end{pmatrix}.$$

The function has Jacobian matrix (derivative), which we write as

$$\dot{S}_\mu(z, x) = \begin{pmatrix} 1 & 0 \\ 0 & \left(\frac{\partial \mu_F(x)}{\partial x} \right)^t \end{pmatrix}.$$

The delta method then gives

$$\begin{aligned} \sqrt{n} \begin{pmatrix} \widehat{\mu}_{\text{np}} - \mu_{\text{true}} \\ \mu_F(\widehat{\theta}_n) - \mu_F(\theta_0) \end{pmatrix} &= \sqrt{n} \begin{pmatrix} \widehat{\mu}_{\text{np}} - \mu_{\text{true}} \\ \widehat{\mu}_{\text{pm}} - \mu_{0,\text{pm}} \end{pmatrix} \\ &\xrightarrow{L} N_2\left(0, (\dot{S}_\mu(\mu_{\text{true}}, \theta_0))^t \Sigma (\dot{S}_\mu(\mu_{\text{true}}, \theta_0))\right) \\ &= N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} V_{\text{np}} & V_{\text{pm,np}} \\ V_{\text{pm,np}}^t & V_{\text{pm}} \end{pmatrix} \right), \end{aligned}$$

which is the limit result that we are proving. What remains now is to validate the first limiting distribution. Using that $\mu_{\text{true}} = \mu(G)$ along with condition (3.3), relation (3.7) may be written as

$$\begin{aligned}\sqrt{n}\begin{pmatrix}\mu(\widehat{G}_n) - \mu_{\text{true}} \\ \widehat{\theta}_n - \theta_0\end{pmatrix} &= \sqrt{n}\begin{pmatrix}\frac{1}{n}\sum_{i=1}^n \text{IF}_\mu(Y_i; G) + o_p\left(\frac{1}{\sqrt{n}}\right) \\ J^{-1}\overline{U}_n + o_p\left(\frac{1}{\sqrt{n}}\right)\end{pmatrix} \\ &= \begin{pmatrix}1 & 0 \\ 0 & J^{-1}\end{pmatrix}\sqrt{n}\begin{pmatrix}\frac{1}{n}\sum_{i=1}^n \text{IF}_\mu(Y_i; G) \\ \frac{1}{n}\sum_{i=1}^n U(Y_i; \theta_0)\end{pmatrix} + \begin{pmatrix}o_p(1) \\ o_p(1)\end{pmatrix}.\end{aligned}$$

Since $U(Y_i; \theta_0)$ has positive definite covariance matrix, the multivariate central limit theorem (B.2.4) gives

$$\sqrt{n}\begin{pmatrix}\frac{1}{n}\sum_{i=1}^n \text{IF}_\mu(Y_i; G) \\ \frac{1}{n}\sum_{i=1}^n U(Y_i; \theta_0)\end{pmatrix} \xrightarrow{L} \begin{pmatrix}\Lambda_{\text{np}} \\ \Lambda_{\text{pm}}\end{pmatrix},$$

where $\Lambda_{\text{np}} \sim N_1(0, \nu)$, and $\Lambda_{\text{pm}} \sim N_p(0, K)$. Slutsky's theorem (B.2.6) consequently gives

$$\sqrt{n}\begin{pmatrix}\mu(\widehat{G}_n) - \mu_{\text{true}} \\ \widehat{\theta}_n - \theta_0\end{pmatrix} \xrightarrow{L} \begin{pmatrix}\Lambda_{\text{np}} \\ J^{-1}\Lambda_{\text{pm}}\end{pmatrix} \overset{d.}{=} N_{p+1}\left(\begin{pmatrix}0 \\ 0\end{pmatrix}, \begin{pmatrix}\Sigma_{00} & \Sigma_{01} \\ \Sigma_{10} & \Sigma_{11}\end{pmatrix}\right),$$

provided the V's take the stated form. This follows since

$$\begin{aligned}\Sigma_{00} &= \text{Var}_G(\Lambda_{\text{np}}) = \nu, \\ \Sigma_{11} &= \text{Var}_G(J^{-1}\Lambda_{\text{pm}}) = J^{-1}\text{Var}_G(\Lambda_{\text{pm}})J^{-1} = J^{-1}KJ^{-1}, \\ \Sigma_{10} &= \Sigma_{01}^t = \text{Cov}_G(\Lambda_{\text{np}}, J^{-1}\Lambda_{\text{pm}}) = J^{-1}\text{Cov}_G(\Lambda_{\text{np}}, \Lambda_{\text{pm}}) = J^{-1}Q.\end{aligned}$$

We have thus completed the proof. □

Now, using that linear transformations of multivariate normal distributions are again univariate normally distributed (theorem B.2.2), this lemma will provide all the limiting distributions we will need later on. The following corollary provides these limiting distributions.

Corollary 3.1.3. *When assumption 3.1.1 holds we get the following limiting distributions*

$$\begin{aligned}\sqrt{n}(\widehat{\mu}_{\text{np}} - \mu_{\text{true}}) &\xrightarrow{L} N\left(0, V_{\text{np}}\right), \\ \sqrt{n}(\widehat{\mu}_{\text{pm}} - \mu_{0,\text{pm}}) &\xrightarrow{L} N_p(0, V_{\text{pm}}), \\ \sqrt{n}(\widehat{b} - b) &\xrightarrow{L} N\left(0, V_{\text{b}}\right),\end{aligned}$$

where

$$\begin{aligned}\widehat{b} &= \widehat{\mu}_{\text{pm}} - \widehat{\mu}_{\text{np}}, \\ b &= \mu_{0,\text{pm}} - \mu_{\text{true}}, \\ V_{\text{b}} &= V_{\text{pm}} + V_{\text{np}} - 2V_{\text{pm,np}}.\end{aligned}$$

Proof. The three relations are all almost direct consequences of lemma 3.1.2. To prove these three results, we apply theorem B.2.2 thrice to the joint limiting distribution of lemma 3.1.2. Specifying the vector a for transformation as $a = (1, 0)$, $a = (0, 1)$ and $a = (1, -1)$ respectively, it is easily seen that the three limiting distributions are exactly what we get. We have thus completed the proof. □

3.2 Mean squared error estimators

In this section we will estimate the mean squared error of $\widehat{\mu}_{\text{np}}$ and $\widehat{\mu}_{\text{pm}}$ by using the limiting distributions we derived in the previous section. The mean squared error is the measure we will use to quantify how good the different estimators are, and based on that we propose a model selection scheme. For an estimator $\widehat{\mu}$, recall that the mse is defined as

$$\text{mse}(\widehat{\mu}) = E_G[(\widehat{\mu} - \mu_{\text{true}})^2] = (E_G[\widehat{\mu}] - \mu_{\text{true}})^2 + \text{Var}_G(\widehat{\mu}) = (\text{bias}(\widehat{\mu}))^2 + \text{Var}_G(\widehat{\mu}).$$

The FIC score will here be defined, in the same manner as in the original FIC apparatus. I.e. the FIC scores forming the criterion will be estimates of the mean squared error:

$$\text{FIC}(\widehat{\mu}) = \widehat{\text{mse}}(\widehat{\mu}) = \widehat{\text{bias}^2}(\widehat{\mu}) + \widehat{\text{Var}}(\widehat{\mu}).$$

Consequently, the estimator (or model) with the smallest value of $\text{FIC}(\widehat{\mu})$ will be chosen by such a scheme. The method for estimating the bias and variance used in this section uses the straightforward empirical analogue of the asymptotic formulae obtained in the previous section. With these asymptotic results, one can get good approximations for the uncertainty of $\widehat{\mu}_{\text{pm}}, \widehat{\mu}_{\text{np}}$ and $\widehat{b}$ for large n. Taking these asymptotic properties as approximations for a sufficiently large n, gives

$$\begin{aligned}
\text{Var}_G(\widehat{\mu}_{\text{np}}) &\approx \frac{1}{n} V_{\text{np}} = \frac{1}{n}\nu, \\
\text{Var}_G(\widehat{\mu}_{\text{pm}}) &\approx \frac{1}{n} V_{\text{pm}} = \frac{1}{n}\left(\left.\frac{\partial \mu_F}{\partial \theta}\right|_{\theta_0}\right)^t J^{-1} K J^{-1} \left(\left.\frac{\partial \mu_F}{\partial \theta}\right|_{\theta_0}\right), \\
\text{Var}_G(\widehat{b}) &\approx \frac{1}{n} V_{\text{b}} = \frac{1}{n}\left(V_{\text{pm}} + V_{\text{np}} - 2V_{\text{pm,np}}\right), \\
&= \frac{1}{n}\left(\left(\left.\frac{\partial \mu_F}{\partial \theta}\right|_{\theta_0}\right)^t J^{-1} K J^{-1} \left(\left.\frac{\partial \mu_F}{\partial \theta}\right|_{\theta_0}\right) + \nu - 2\left(\left.\frac{\partial \mu_F}{\partial \theta}\right|_{\theta_0}\right)^t J^{-1} Q\right).
\end{aligned}$$

Thus, straightforward reasonable estimators for these variances may be obtained by simply taking the empirical analogues of these formulae. Doing that leads to the following estimators:

$$\widehat{V}_{\text{np}} = \widehat{\nu}, \tag{3.8}$$

$$\widehat{V}_{\text{pm}} = \left(\left.\frac{\partial \mu_F}{\partial \theta}\right|_{\widehat{\theta}_n}\right)^t \widehat{J}^{-1} \widehat{K} \widehat{J}^{-1} \left(\left.\frac{\partial \mu_F}{\partial \theta}\right|_{\widehat{\theta}_n}\right), \tag{3.9}$$

$$\widehat{V}_{\text{pm,np}} = \left(\left.\frac{\partial \mu_F}{\partial \theta}\right|_{\widehat{\theta}_n}\right)^t \widehat{J}^{-1} \widehat{Q},$$

$$\widehat{V}_{\text{b}} = \widehat{V}_{\text{pm}} + \widehat{V}_{\text{np}} - 2\widehat{V}_{\text{pm,np}}, \tag{3.10}$$

where

$$\widehat{J} = -\frac{1}{n}\sum_{i=1}^{n} I(Y_i, \widehat{\theta}_n),$$
$$\widehat{K} = \frac{1}{n}\sum_{i=1}^{n} U(Y_i;\widehat{\theta}_n)U(Y_i;\widehat{\theta}_n)^t,$$
$$\widehat{\nu} = \frac{1}{n}\sum_{i=1}^{n}\Big(\mathrm{IF}_\mu(Y_i;\widehat{G}_n)\Big)^2 = \frac{1}{n}\sum_{i=1}^{n}\mathrm{IF}_\mu(Y_i;\widehat{G}_n)^2,$$
$$\widehat{Q} = \frac{1}{n}\sum_{i=1}^{n} U(Y_i;\widehat{\theta}_n)\mathrm{IF}_\mu(Y_i;\widehat{G}_n).$$

We are now ready to give estimators for the mse of $\widehat{\mu}_{\mathrm{np}}$ and $\widehat{\mu}_{\mathrm{pm}}$.

3.2.1 Estimator in the nonparametric case

From corollary 3.1.3 we see that the expectation of $\sqrt{n}(\widehat{\mu}_{\mathrm{np}} - \mu_{\mathrm{true}})$ converge to zero. It is therefore reasonable to estimate the bias of the nonparametric estimator by zero:[1]

$$\widehat{\mathrm{bias}}_{\mathrm{np}} = 0.$$

As a consequence of this, the natural estimate of the squared bias of $\widehat{\mu}_{\mathrm{np}}$ is also zero:

$$\widehat{\mathrm{bias}^2}_{\mathrm{np}} = 0.$$

The corresponding estimate of the variance of $\widehat{\mu}_{\mathrm{np}}$ is already found in equation (3.8) to be

$$\widehat{\mathrm{Var}}(\widehat{\mu}_{\mathrm{np}}) = \frac{1}{n}\widehat{V}_{\mathrm{np}} = \frac{1}{n}\widehat{\nu}.$$

The straightforward empirical estimator for $\mathrm{mse}(\widehat{\mu}_{\mathrm{np}})$ is therefore given by

$$\widehat{\mathrm{mse}}(\widehat{\mu}_{\mathrm{np}}) = \widehat{\mathrm{bias}^2}_{\mathrm{np}} + \widehat{\mathrm{Var}}(\widehat{\mu}_{\mathrm{np}}) = \frac{1}{n}\widehat{V}_{\mathrm{np}}.$$

3.2.2 Estimator in the parametric case

In the parametric case, observe first that equation (3.9) gives a natural estimator of the variance involved for the parametric estimator. Therefore

$$\widehat{\mathrm{Var}}(\widehat{\mu}_{\mathrm{pm}}) = \frac{1}{n}\widehat{V}_{\mathrm{pm}},$$

is considered a good estimator of the variance of $\widehat{\mu}_{\mathrm{pm}}$. From corollary 3.1.3 we also see that the expectation of $\sqrt{n}(\widehat{\mu}_{\mathrm{pm}} - \mu_{0,\mathrm{pm}})$ converges to 0. Since $\mu_{0,\mathrm{pm}}$ and μ_{true} does not vary with n, this implies that when $\mu_{0,\mathrm{pm}} \neq \mu_{\mathrm{true}}$ the parametric estimator $\widehat{\mu}_{\mathrm{pm}}$ does *not* have the property of being asymptotically unbiased. The estimator of the squared bias of $\widehat{\mu}_{\mathrm{pm}}$ is therefore crucial

[1] In many situations the expectation is exactly zero also for finite n, or a simple modification may be applied to achieve this without changing the asymptotics.

in this case. When deriving an estimate of the squared bias of $\widehat{\mu}_{\text{pm}}$, it is natural to start with an estimator of the (non-squared) bias:

$$\text{bias}_{\text{pm}} \stackrel{\text{def.}}{=} E_G[\widehat{\mu}_{\text{pm}} - \mu_{\text{true}}] = E_G[\widehat{\mu}_{\text{pm}} - \mu_{0,\text{pm}}] + b.$$

Since both $\sqrt{n}(\widehat{\mu}_{\text{pm}} - \mu_{0,\text{pm}})$ and $\sqrt{n}(\widehat{b} - b)$ converge in law to zero-mean random variables by corollary 3.1.3, we have by theorem B.2.3 that $\widehat{\mu}_{\text{pm}} \xrightarrow{P} \mu_{0,\text{pm}}$ and $\widehat{b} \xrightarrow{P} b$. It is thus natural to estimate $E_G[\widehat{\mu}_{\text{pm}} - \mu_{0,\text{pm}}]$ by 0 and b by $\widehat{b}$. This leads to

$$\widehat{\text{bias}}_{\text{pm}} = 0 + \widehat{b} = \widehat{\mu}_{\text{pm}} - \widehat{\mu}_{\text{np}} = \mu_F(\widehat{\theta}_n) - \mu(\widehat{G}_n).$$

The main interest is however on the squared bias. The straightforward approach is perhaps to use the estimator

$$\widehat{\text{bias}^2}^*_{\text{pm}} = (\widehat{\text{bias}}_{\text{pm}})^2 = (\widehat{\mu}_{\text{pm}} - \widehat{\mu}_{\text{np}})^2 = \left(\widehat{b}\right)^2, \tag{3.11}$$

but since $b = \mu_{0,\text{pm}} - \mu_{\text{true}} \approx E_G[\widehat{\mu}_{\text{pm}} - \mu_{\text{true}}] = \text{bias}_{\text{pm}}$, and

$$\begin{aligned} E_G[\widehat{\text{bias}^2}^*_{\text{pm}}] = E_G[(\widehat{b})^2] &= \left(E_G[\widehat{b}]\right)^2 + \text{Var}_G(\widehat{b}) \\ &\approx b^2 + \text{Var}_G(\widehat{b}) \approx (\text{bias}_{\text{pm}})^2 + \text{Var}_G(\widehat{b}), \end{aligned}$$

such an estimator will in general overestimate the intended squared bias because the variance of the bias estimator is nonzero for finite n. As seen from the above formulae, an approximately unbiased estimator of the squared bias may be formed by subtracting an estimate of the additional variance. From corollary 3.1.3 and equation (3.10) we have that $\text{Var}_G(\widehat{b})$ may be estimated by $\frac{1}{n}\widehat{V}_{\text{b}}$. Hence, adjusting formula (3.11) using this variance estimator, gives

$$\begin{aligned} \widehat{\text{bias}^2}_{\text{pm}} &= \widehat{\text{bias}^2}^*_{\text{pm}} - \frac{1}{n}\widehat{V}_{\text{b}} \\ &= (\widehat{\mu}_{\text{pm}} - \widehat{\mu}_{\text{np}})^2 - \frac{1}{n}\left(\widehat{V}_{\text{pm}} + \widehat{V}_{\text{np}} - 2\widehat{V}_{\text{pm,np}}\right). \end{aligned}$$

In total we then get an estimator of the mse given by

$$\begin{aligned} \widehat{\text{mse}}(\widehat{\mu}_{\text{pm}}) &= \widehat{\text{bias}^2}_{\text{pm}} + \widehat{\text{Var}}(\widehat{\mu}_{\text{pm}}) \\ &= (\widehat{\mu}_{\text{pm}} - \widehat{\mu}_{\text{np}})^2 - \frac{1}{n}\left(\widehat{V}_{\text{pm}} + \widehat{V}_{\text{np}} - 2\widehat{V}_{\text{pm,np}}\right) + \frac{1}{n}\widehat{V}_{\text{pm}} \\ &= (\widehat{\mu}_{\text{pm}} - \widehat{\mu}_{\text{np}})^2 - \frac{1}{n}\widehat{V}_{\text{np}} + 2\frac{1}{n}\widehat{V}_{\text{pm,np}}. \end{aligned}$$

Note that the estimator of the variance of $\widehat{\mu}_{\text{pm}}$ disappears and simplifies the equation. Now, let the FIC scores be given by

$$\text{FIC}(\widehat{\mu}_{\text{np}}) = \widehat{\text{mse}}(\widehat{\mu}_{\text{np}}) = \frac{1}{n}\widehat{V}_{\text{np}}, \tag{3.12}$$

$$\text{FIC}(\widehat{\mu}_{\text{pm}}) = \widehat{\text{mse}}(\widehat{\mu}_{\text{pm}}) = (\widehat{\mu}_{\text{pm}} - \widehat{\mu}_{\text{np}})^2 - \frac{1}{n}\widehat{V}_{\text{np}} + 2\frac{1}{n}\widehat{V}_{\text{pm,np}}, \tag{3.13}$$

and define a model selection scheme as follows: The model whose estimator has the smallest value according to the FIC scores given by formulae (3.12) and (3.13) are selected among the possibly k different parametric models and the nonparametric model. We have thus established a FIC framework that may be used for focused model selection when the estimators are based on nonparametric and parametric plug-in estimators of the focus parameter μ. As mentioned in the introduction of this thesis, the above formuale applies not only to situations with one single parametric model, but clearly also to situations with several parametric models.

3.2.3 An adjusted FIC scheme

Since the final FIC scores of equations (3.12) and (3.13) are based on asymptotic properties of other estimators, and then approximated only by the use of the finite number of data points available, there is no guarantee that estimates produced for every model are reasonable for small n.[2] For the FIC score to be reasonable the estimates should have the property that any squared bias and variance estimate is nonnegative. In the nonparametric case this is already fulfilled since the squared bias estimate is zero and the variance consists of squared functions of the data. In the parametric case, however, this is not always the case. A good estimator should fulfill the following inequalities:

- $\widehat{\text{bias}^2}_{\text{pm}} \geq 0 \Leftrightarrow (\widehat{\mu}_{\text{pm}} - \widehat{\mu}_{\text{np}})^2 \geq \frac{1}{n}\widehat{V}_{\text{b}} = \frac{1}{n}\left(\widehat{V}_{\text{pm}} + \widehat{V}_{\text{np}} - 2\widehat{V}_{\text{pm,np}}\right)$,
- $\widehat{V}_{\text{b}} \geq 0 \Leftrightarrow \widehat{V}_{\text{pm}} + \widehat{V}_{\text{np}} \geq 2\widehat{V}_{\text{pm,np}}$.

If these restrictions are not met, the estimates are individually not meaningful. Observe that to check these inequalities, we have to calculate $\widehat{V}_{\text{pm}}$ even if it is not included in the FIC formula in equation (3.13). To overcome the possible problems when these restrictions are not met, we propose an adjusted FIC apparatus by setting the critical quantities of equation (3.13) to zero if their estimated values are negative. Doing so leads to the following new formula for $\text{FIC}(\widehat{\mu}_{\text{pm}})$:

$$\begin{aligned}\text{FIC}^*(\widehat{\mu}_{\text{pm}}) = &\left\{(\widehat{\mu}_{\text{pm}} - \widehat{\mu}_{\text{np}})^2 - \left[\widehat{\text{Var}}(\widehat{\mu}_{\text{pm}}) + \widehat{\text{Var}}(\widehat{\mu}_{\text{np}}) - 2\widehat{\text{Cov}}(\widehat{\mu}_{\text{pm}}, \widehat{\mu}_{\text{np}})\right]^+\right\}^+ \\ &+ \widehat{\text{Var}}(\widehat{\mu}_{\text{pm}}),\end{aligned}$$

where

$$[z]^+ = \{z\}^+ = \begin{cases} z & , z > 0 \\ 0 & , \text{otherwise.} \end{cases}$$

The expression is slightly more complicated than the original, but as mentioned it has the advantage that it produces reasonable estimates for all terms of the mse formulae. It is therefore reasonable to apply this formula to all model selection problems of this types, regardless of the numerical values produced in the estimation process. Replacing $\text{FIC}(\widehat{\mu}_{\text{pm}})$ by $\text{FIC}^*(\widehat{\mu}_{\text{pm}})$ while $\text{FIC}(\widehat{\mu}_{\text{np}})$ remains the same, gives in this sense a more robust model selection scheme.

[2] What "small" means varies from situation to situation.

It should also be noted that in most of the natural situations, the covariance between the nonparametric estimator and each of the parametric estimators should be positively correlated. That is

$$\widehat{\mathrm{Cov}}(\widehat{\mu}_{\mathrm{pm}}, \widehat{\mu}_{\mathrm{np}}) > 0. \tag{3.14}$$

Since this is not an actual restriction, we do not require this specifically, but it should nevertheless be checked to make sure everything has been done correctly. There should really be no situations except possibly constructed special cases where inequality (3.14) does not hold, at least when the parametric family one is working with is not totally wrong.

3.2.4 Numerical approximations

The two derived schemes rely heavily on the functions of the parametric distribution. To apply the formulae above, it is required that one possesses analytical expressions for the score function and the information function in addition to the derivative of the focus parameter with respect to the model parameters. All these quantities involve differentiation with respect to the model parameters. Although they are fairly easily obtained in many cases, they are hard find or even impossible to express analytically for some parametric models. To circumvent these problems we suggest using numerical approximations for these quantities. The usual approximation to the derivative of a function $S(x)$ is

$$\dot{S}_{\mathrm{approx}}(x) = \frac{S(x+\epsilon) - S(x-\epsilon)}{2\epsilon},$$

for a small number ϵ (say $\epsilon = 10^{-6}$). This method may work fairly well in many cases, but it is numerically unstable. There does however exist more complicated methods to handle numerical derivation which are included in most programming languages and software packages used for statistics.

Although integrals are not directly represented in the FIC formulae presented above, one may nevertheless have to solve some integrals on the way to the FIC formulae. Mostly this concerns the actual calculation of the focus parameter estimate under the different models. As for derivation, this is a straightforward task in some situations, while it is more difficult in others, which calls for numerical integration techniques. Both these themes will be discussed slightly further in chapter 8.

3.3 Sufficient conditions and concrete situations

In the previous section we derived FIC schemes based on a limiting distribution derived under a few assumptions. In this section we will investigate what kind of situations that meet the stated requirements. We will discuss conditions implying some of the key ingredients of assumption 3.1.1, and give examples of situations that may be handled in this framework in addition to pointing out a few situations that cannot be handled by the developed apparatus.

As a heads-up we note that when we in the following will be working with Hadamard and Fréchet differentiability, we will use the norm representation and not the metric representation. This is done for consistency with other representations in this thesis, and also because all metrics of interest are based on norms.

3.3.1 Statements and discussion of sufficient conditions

The following lemma states useful conditions for the focus parameter to be "smooth" enough to be handled by the proposed FIC apparatus.

Lemma 3.3.1. *Each of the following two conditions are sufficient for condition* (3.1) *to hold:*

(i) μ *is Hadamard differentiable at* G *for each* $y \in \Omega$*, with respect to the supremum norm* $\|\cdot\|_\infty$.

(ii) μ *is Fréchet differentiable at* G *for each* $y \in \Omega$ *with a norm* $\|\cdot\|_*$ *satisfying*

$$\sqrt{n}\|\widehat{G}_n - G\|_* = O_p(1). \tag{3.15}$$

Proof. (i) The proof is given in Fernholz (1983).
(ii) Let

$$K_n = (\mu(\widehat{G}_n) - \mu(G)) - \frac{1}{n}\sum_{i=1}^{n} \mathrm{IF}_\mu(Y_i; G).$$

From the definition of Fréchet differentiability (definition B.1.1, iii) we have that for any $\epsilon > 0$ there exist a $\delta > 0$ such that $|K_n| < \epsilon\|\widehat{G}_n - G\|_*$, whenever $\|\widehat{G}_n - G\|_* < \delta$. Thus we have that

$$\begin{aligned} &\|\widehat{G}_n - G\|_* < \delta \Rightarrow |K_n| < \epsilon\|\widehat{G}_n - G\|_*, \\ &\qquad\qquad \Updownarrow \\ &\|\widehat{G}_n - G\|_* > \delta \Rightarrow |K_n| > \epsilon\|\widehat{G}_n - G\|_*, \end{aligned} \tag{3.16}$$

which implies that

$$Pr\left\{\|\widehat{G}_n - G\|_* > \delta\right\} \geq Pr\left\{|K_n| > \epsilon\|\widehat{G}_n - G\|_*\right\}.$$

Now, using standard results of probability theory, we get

$$\begin{aligned} Pr\left\{\|\widehat{G}_n - G\|_* > \delta\right\} &\geq Pr\left\{\sqrt{n}|K_n| > \epsilon\sqrt{n}\|\widehat{G}_n - G\|_*\right\} \\ &\geq Pr\left\{\sqrt{n}|K_n| > \eta \geq \epsilon\sqrt{n}\|\widehat{G}_n - G\|_*\right\} \\ &= Pr\left\{\sqrt{n}|K_n| > \eta \cap \eta \geq \epsilon\sqrt{n}\|\widehat{G}_n - G\|_*\right\} \\ &= Pr\left\{\sqrt{n}|K_n| > \eta\right\} + Pr\left\{\eta \geq \epsilon\sqrt{n}\|\widehat{G}_n - G\|_*\right\} \\ &\quad - Pr\left\{\sqrt{n}|K_n| > \eta \cup \eta \geq \epsilon\sqrt{n}\|\widehat{G}_n - G\|_*\right\} \\ &\geq Pr\left\{\sqrt{n}|K_n| > \eta\right\} + Pr\left\{\eta \geq \epsilon\sqrt{n}\|\widehat{G}_n - G\|_*\right\} - 1 \\ &= Pr\left\{\sqrt{n}|K_n| > \eta\right\} - Pr\left\{\eta < \epsilon\sqrt{n}\|\widehat{G}_n - G\|_*\right\}. \end{aligned}$$

Hence,

$$Pr\left\{\sqrt{n}|K_n| > \eta\right\} \leq Pr\left\{\|\widehat{G}_n - G\|_* > \delta\right\} + Pr\left\{\eta/\epsilon < \sqrt{n}\|\widehat{G}_n - G\|_*\right\}.$$

Taking the limit as $n \to \infty$ on both sides of the inequality gives

$$\begin{aligned}\lim_{n\to\infty} Pr\left\{\sqrt{n}|K_n| > \eta\right\} &\leq \lim_{n\to\infty} Pr\left\{\|\widehat{G}_n - G\|_* > \delta\right\} + \lim_{n\to\infty} Pr\left\{\eta/\epsilon < \sqrt{n}\|\widehat{G}_n - G\|_*\right\} \\ &= \lim_{n\to\infty} Pr\left\{\eta/\epsilon < \sqrt{n}\|\widehat{G}_n - G\|_*\right\},\end{aligned}$$

since $\lim_{n\to\infty} Pr\left\{\|\widehat{G}_n - G\|_* > \delta\right\} = 0$ for any δ by the assumption. Now, since ϵ can be chosen arbitrarily small it follows from condition (3.15) that the right hand side is zero for some sufficiently small ϵ, which furthermore implies that the left hand side also is zero. Then it follows from the definition of convergence in probability that $K_n = o_p(1/\sqrt{n})$, which completes the proof. □

Since Hadamard differentiability is a weaker form of differentiability than that of Fréchet, where the latter implies the former for the same metric, one might think that condition (ii) above is rather pointless. Note however that condition (ii) is not restricted to the supremum norm. There are in fact cases where the functional is not Hadamard differentiable with respect to some norm, but where the same functional is Fréchet differentiable with respect to different norm. In fact, for the class of functionals $T(H) = \phi\left(\int x dH(x)\right)$, where ϕ is a differentiable function $\phi : \mathbb{R}^m \mapsto \mathbb{R}$, there exists cases where T is not Hadamard differentiable with respect to the supremum norm, while it is Fréchet differentiable with respect to the L_1-norm given by: $\rho(H_1, H_2) = \int |H_1(s) - H_2(s)|\, \mathrm{d}s$, see Shao (2003, p. 340). The following corollary is useful in some situations where μ is not Hadamard differentiable with respect to the supremum norm.

Corollary 3.3.2. *Let $\|\cdot\|_* = \|\cdot\|_p$ denote the L_p-norm (or simply p-norm) , i.e.*

$$\|H_1 - H_2\|_p = \left[\int |H_1(s) - H_2(s)|^p \, \mathrm{d}s\right]^{1/p},$$

with $p \geq 1$. When either $p > 2$, or the two conditions $1 \leq p < 2$ and $\int [G(s)(1 - G(s))]^{p/2} ds < \infty$ are satisfied, then the norm $\|\cdot\|_* = \|\cdot\|_p$ satisfies condition (3.15) of lemma 3.3.1 (ii).

Proof. The result follows if one is able to show that $E_G[\|\widehat{G}_n - G\|] = O(1)$, since if the sequence of expectations of random variables is bounded, the sequence of random variables is also bounded in probability. However, that $E_G[\|\widehat{G}_n - G\|] = O(1)$ is shown using the two famous inequalities statements of Hölder and Jensen, as given in Shao (2003, proof of Theorem 5.2 (ii)). □

The conditions and lemmas of this section have so far concerned the nonparametrics and especially equation (3.1). Turning to the parametrics and equation (3.3), one may state conditions of a more traditional form. We could have stated quite similar conditions also for the parametrics by treating θ as a functional of a cdf, but since it is more common to work just with the quantities θ_0, J and $\overline{U}_n$, we follow this practice as well. The following theorem provides precise and quite simple sufficient conditions under which relation (3.3) holds.

Theorem 3.3.3. (Asymptotic normality of ML estimators, rewritten from van der Vaart (2000, theorem 5.41))
Let $Y_1, \ldots, Y_n$ be iid random variables from an unknown distribution with cdf $G(y)$. Let $f(y;\theta)$ be the density or pmf of a class of parametric distributions with p-dimensional parameter vector θ. Suppose also that the following conditions hold:

- *$\widehat{\theta}_n$ is the only root of $\overline{U}_n(\theta)$ for every n large enough.*
- *θ_0 is an interior point of the parameter space Θ.*
- *The score function $U(y;\theta)$ is twice continuously differentiable in θ for every y.*
- *J exists, and is nonsingular.*
- *The second order partial derivatives of the score function $U(y;\theta)$ with respect to θ, are dominated by a fixed integrable function $K_0(y)$ for every θ in a neighborhood of θ_0.*

Under these conditions, the relation (3.3) *of assumption 3.1.1 holds.*

Proof. The proof is analogous to the proof of van der Vaart (2000, theorem 5.41). It consists of arguments including a Taylor expansion, the central limit theorem (B.2.4) and careful use of the assumptions to make sure the functions involved does not behave unsatisfactory. The theorem does however assume that $\widehat{\theta}_n$ is consistent for θ_0, but as van der Vaart (2000, theorem 5.42) shows that $\widehat{\theta}_n$ is consistent when it is the only root of $\overline{U}_n(\theta)$, the result follows. □

Sufficient conditions for the two key quantities of assumption 3.1.1 have now been given. The statements in condition (3.2) can often be checked by direct computation, but the remaining conditions are in most situations impossible to check directly without additional information of the unknown features of the data. One therefore simply has to assume they are properties of the unknown distribution. The statements in condition (3.4) involve both unknown least false parameters and the unknown true distribution and are therefore hard to check. Still, the conditions are rather weak and will hold in most situations of practical interest. It is therefore reasonable to assume these properties. Condition (3.5) makes sure the delta method works properly and returns a variable with positive variance. The condition is included more of less for completeness. If one defines $N(0,0)$ as the scalar zero, the condition is not needed for the results to hold. The uniqueness assumption of θ_0 is also weak. There exist conditions including log-concavity of $f(y;\theta)$, which assures that there is a unique minimizer of the Kullback–Leibler divergence, but it is more common to just assume such a property. That the focus parameter is one-dimensional is also a condition included in assumption 3.1.1. This assumption may at first sight seem a bit peculiar, but rephrasing the FIC idea should clarify it all. The FIC idea is to "choose the best model for estimating a certain focus parameter". Choosing the model that are best at estimating a multidimensional focus parameter will be equivalent to choosing the model that is best at more than one estimation task. This is not our intention and therefore the focus parameter is restricted to one dimension. A model selection routine choosing the model that is overall best at estimating several parameters will be the theme of chapter 6.

3.3.2 Applicable and non-applicable situations

From definition B.1.1 of Fréchet differentiability it is clear that all focus parameters μ which may be written as a linear functional is Fréchet differentiable for any norm. This implies that all focus parameters that can be written as $\mu(H) = \int S(z)\,\mathrm{d}H(z)$, where S is an integrable function $s : \mathbb{R} \mapsto \mathbb{R}$, fulfill condition (3.1). For the slightly more general functionals $\mu(H) = \phi\left(\int S(z)\,\mathrm{d}H(z)\right)$, where S is an integrable function $S : \mathbb{R} \mapsto \mathbb{R}^k$ and ϕ is a differentiable function $S : \mathbb{R}^k \mapsto \mathbb{R}$, we cannot draw such a conclusion directly from the definition. As a matter of fact Shao (2003, chapter 5.2) states that not all functionals, even where $S(z) = z$, are

Hadamard differentiable with respect to the supremum norm. Such functionals do however turn out to be Fréchet differentiable with respect to the L_1-norm defined above, and by corollary 3.3.2, the functionals apply under the additional moment condition of the corollary.

Going via lemma 3.3.1 is one way of checking that the focus parameter can be applied to the current situation and condition (3.1) holds. Another strategy is to check condition (3.1) directly. Consider now the set of focus parameters which can be represented as

$$\mu(H) = \phi\left(\int S(z)\,\mathrm{d}H(z)\right), \tag{3.17}$$

where S is an integrable function $S : \mathbb{R} \mapsto \mathbb{R}^k$ and ϕ is smooth function $S : \mathbb{R}^k \mapsto \mathbb{R}$.[3] We will refer to this type of focus parameter as a functional of the smooth function of averages class. This class will be central in what follows. Note that most results stated for this class of functionals remains true also when ϕ is only differentiable once. The class of smooth functions of averages is a very wide and important class of focus parameters which includes some of the most natural focus parameters. Among them are all moments, the cdf and differentiable functions of these. As a result, the mean, variance, standard deviation, skewness, kurtosis and the probability that a specific event occurs, are all of this type. The influence function for a functional of this class is given by

$$\begin{aligned}\mathrm{IF}_\mu(y;H) &= \frac{\partial}{\partial\lambda}\phi\left(\int S(z)\,\mathrm{d}(H+\lambda(\delta_y - H))(z)\right)\Big|_{\lambda=0} \\ &= \dot\phi\left(\int S(z)\,\mathrm{d}H(z)\right)^t\left(S(y) - \int S(z)dH(z)\right),\end{aligned} \tag{3.18}$$

where $\dot\phi$ denotes the gradient of ϕ, i.e. the column vector of partial derivatives of ϕ. Note that $S(y)$ here is a simplified form of $(S_1(y),\ldots,S_k(y))^t$. As a result we get

$$\overline{\mathrm{IF}}_{\mu,n}(G) = \dot\phi\left(\int S(z)\,\mathrm{d}G(z)\right)\left(\frac{1}{n}\sum_{i=1}^n S(Y_i) - \int S(z)dG(z)\right).$$

However, a regular Taylor expansion (theorem B.2.7) of $\phi(a)$ evaluated at b gives

$$\phi(a) = \phi(b) + \dot\phi(b)(a-b) + o(\|a-b\|).$$

Note that $\mu(\widehat{G}_n) = \phi(a)$ and $\mu(G) = \phi(b)$. By using this fact and letting $a = (1/n)\sum_{i=1}^n S(Y_i)$ and $b = \int S(z)dH(z)$, we get

$$\mu(\widehat{G}_n) = \mu(G) + \overline{\mathrm{IF}}_{\mu,n}(G) + o_p\left(\left\|\frac{1}{n}\sum_{i=1}^n S(Y_i) - \int S(z)dG(z)\right\|\right).$$

However,$(1/n)\sum_{i=1}^n S(Y_i) - \int S(z)dG(z)$ is simply the expectation of $S(Y_i)$ subtracted by the mean of these iid variables. By the central limit theorem (B.2.4), we get that $\sqrt{n}$ times this factor converges in distribution provided $E_G[\|S(Y_i)\|^2] < \infty$. Thus, under this assumption we get that

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n S(Y_i) - \int S(z)dG(z)\right) = O_p(1),$$

[3] A smooth function is a function that is continuously differentiable an infinite number of times.

which furthermore gives

$$o_p\left(\left\|\frac{1}{n}\sum_{i=1}^{n} S(Y_i) - \int S(z)dG(z)\right\|\right) = o_p\left(O_p\left(\frac{1}{\sqrt{n}}\right)\right) = o_p\left(\frac{1}{\sqrt{n}}\right).$$

However, the additional assumption of $E_G[\|S(Y_i)\|^2] < \infty$ is equivalent to the assumption that $E_G\left[\mathrm{IF}_\mu(Y_i;G)^2\right] < \infty$, which already is a condition in assumption 3.1.1. As a result, all focus parameters that may be written on the form of equation (3.17), fulfills condition (3.1) whenever the second statement in condition (3.2) is also fulfilled. The first statement in condition (3.2) will always be fulfilled for this class.

As a fairly easy example, consider estimation of the pmf at some point y_0 for a discrete distribution. For a discrete distribution with cdf H and pmf h, the focus parameter $h(y_0)$ may be written as $\mu(H) = Pr_H\{Y_i = y_0\}$. By rewriting this expression we also see that it can be written as an integral with respect to H, which is also on the smooth form above. This follows since

$$\mu(H) = Pr_H\{Y_i = y_0\} = h(y_0) = \sum_{y\in\Omega_H} \mathbf{1}_{\{y=y_0\}}(y)h(y) = \int \mathbf{1}_{\{y=y_0\}}(y)\,\mathrm{d}H(y), \tag{3.19}$$

for Ω the discrete sample space of Y_i. Plugging in the ecdf, we get

$$\widehat{\mu}_{\mathrm{np}} = \mu(\widehat{G}_n) = \int \mathbf{1}_{\{x=y_0\}}(x)\,\mathrm{d}\widehat{G}_n(x) = \frac{\#\{Y_i = y_0\}}{n},$$

which is just the proportion of the data that equals y_0. Since the focus parameter is on the smooth function of averages form it satisfies condition (3.1) whenever condition two of (3.2) also is fulfilled. By formula (3.18), we get that

$$\mathrm{IF}_\mu(Y_i;G) = \mathbf{1}_{\{Y_i=y_0\}}(Y_i) - \mu(G).$$

Thus, we see that

$$\begin{aligned} E_G\left[\left(\mathbf{1}_{\{Y_i=y_0\}}(Y_i) - g(y_0)\right)^2\right] &= g(y_0) - 2g(y_0)^2 + g(y_0)^2 \\ &= g(y_0)(1-g(y_0)) \le 1/4 < \infty. \end{aligned}$$

Hence, the second statement in condition (3.2) is fulfilled for any discrete true distribution with pmf g, which shows that both the conditions (3.1) and (3.2) are fulfilled for this useful focus parameter. It is however of great importance that the true distribution is discrete. The reason for this will be clear from the discussion below.

The whole scheme is built on the assumption that the nonparametric estimator is just the plug-in estimator $\widehat{\mu}_{\mathrm{np}} = \mu(\widehat{G}_n)$. In most cases this is a good nonparametric estimator. Nevertheless, there exists focus parameters which is not very well estimated by just plugging in the ecdf. Consider the density of a continuous distribution at some point y_0. This focus parameter may as a functional be written as

$$\mu(H) = \left.\frac{\partial H(y)}{\partial y}\right|_{y=y_0}.$$

This is a linear functional so equation (3.1) should hold. The nonparametric plug-in estimator is nonetheless given by

$$\mu(\widehat{G}_n) = \left.\frac{\partial \widehat{G}_n(y)}{\partial y}\right|_{y=y_0} = 0,$$

when there is no $Y_i = y_0$, and is undefined when there is at least one $Y_i = y_0$. By redefining the density as the right derivative of the cdf, the estimator is always defined, but it will still always return zero. Even if this is a meaningless estimator, some "crazy mind" could possibly consider going on after all. However, some algebra shows that the influence function turns out to be $\text{IF}_\mu(y; G) = -g(y_0)$ for every y. Hence the statements in condition (3.2) are not fulfilled, and the scheme cannot be applied. Density estimation based on nonparametrics is however better estimated using kernel functions. In section 5.1 of chapter 5, we discuss FICology for this type of nonparametric density estimation.

Even if the focus parameter class above is a fairly wide class, it does not span over all types of focus parameters one could possibly be interested in. Especially, the quantile function $\mu(H) = H^{-1}(y_0)$ for a continuous distribution evaluated at some point y_0, is not of this type. The median is certainly a special case of the quantile function where $y_0 = 1/2$. This type of focus parameter is however seen to be Hadamard differentiable with respect to the supremum norm (van der Vaart (2000, lemma 21.3)), and as a result of lemma 3.3.1 (i), the quantile function may also be applied to the scheme. A fairly useful property of Hadamard differentiability is that it possesses a chain rule. For the precise statement of the chain rule, see van der Vaart (2000, theorem 20.9).

In addition to regularity of the focus parameters, assumption 3.1.1 lays restrictions on the parametric models and the true underlying distribution. The following illustration shows that one does not have to invent something extraordinary for the assumptions not to fit. Consider the situation where the median is of interest and the true but unknown distribution is the Cauchy distribution. The median is perfectly defined for the Cauchy distribution, so this is certainly a valid situation. However, if the normal distribution is among the competing models, condition (3.4) would not be satisfied, simply because the score function would not have a finite expectation. Thus, the normal distribution spoils the fun in this situation. If one chooses a different competing parametric distribution, it may turn out fine, but such a distribution cannot have a score function where moments are represented, which is rather unpleasant. The reason for this is that no moment of the Cauchy distribution exists. Note also that if the original focus parameter was some moment, no parametric distribution could make the situation work out since the nonparametric estimator will eventually run away. Summing up, these problems are all caused by the troublesome Cauchy distribution. Fortunately the Cauchy distribution is rarely represented in nature.

As a final illustration, consider the simple case where the Uniform distribution $U[0, \theta]$ is among the set of competing models. By some mathematics (see e.g. Lehmann (1998, Example 2.3.7)), one gets that $\widehat{\theta}_n = \max_{i=1,\ldots,n}(Y_i)$ which converges faster than the usual $\sqrt{n}$-rate. In fact it can be shown that $n(\widehat{\theta}_n - \theta_0) \xrightarrow{L} \text{Exp}(1/\theta_0)$. Thus, condition (3.3) is not fulfilled, and the theory does not hold for this situation. This particular situation will be discussed further in section 5.6 of chapter 5.

Finally we note that even if we have illustrated a few cases where the assumptions does not hold, assumption 3.1.1 is weak and most practical situations one would encounter with real data, works perfectly fine.

3.4 Consistency and unbiasedness of FIC scheme estimators

An optimal estimator possesses the property of being both consistent and unbiased. Being in possession of both these properties are however quite rare and are mostly reserved for the simplest estimators. The main interest usually lies in the properties of the final estimators, which here corresponds to the mse estimators. Since these estimators (at least in the parametric case) consists of terms of different convergence rates, a global treatment of the mse estimators are not very accommodating. It is more interesting to investigate the properties when each of individually estimators are stabilized.

In this section we deal with these properties for the estimators included in the FIC scheme presented in section 3.2. With the splitting explained above, we will especially show that all variance estimators are strongly consistent, and that the same holds in an asymptotic sense for the squared bias estimators. Finally we will discuss the bias estimator in greater detail.

3.4.1 Consistency for the focus parameter estimators

Since both $\widehat{\mu}_{\text{np}}$ and $\widehat{\mu}_{\text{pm}}$ are included in the parametric FIC formula, we would like to show that both of them are consistent estimators for their respective estimands μ_{true} and $\mu_{0,\text{pm}}$. Since these are the estimators the whole FIC scheme is built on, consistency of these is of high separate interest since they indicate that the estimation process is in some sense "good". The following lemma gives quite weak conditions where we are able to prove strong consistency for both the parametric and nonparametric μ estimators.

Lemma 3.4.1. *Let $\mu(G)$ be continuous as a functional in the cdf G with respect to the supremum norm $\|\cdot\|_\infty$, $\mu_F(\theta)$ is continuous in θ_0 and the regularity conditions of theorem 3.3.3 hold. Then $\widehat{\mu}_{\text{np}}$ is a strongly consistent estimator for μ_{true} and $\widehat{\mu}_{\text{pm}}$ is a consistent estimator for $\mu_{0,\text{pm}}$.*

Proof. By the Glivenko–Cantelli theorem (B.2.10), $\|\widehat{G}_n - G\|_\infty \xrightarrow{a.s.} 0$. By the continuous mapping theorem (B.2.9), it follows that $\widehat{\mu}_{\text{np}} \xrightarrow{a.s.} \mu_{\text{true}}$. By the arguments in the proof of theorem 3.3.3, we get that $\widehat{\theta}_n \xrightarrow{P} \theta_0$. By the continuous mapping theorem (B.2.9) it thus follows that $\widehat{\mu}_{\text{pm}} \xrightarrow{P} \mu_{0,\text{pm}}$. □

Note that strong consistency for the ML estimator may be established as well. See e.g. Huber (1967, case A).

3.4.2 Consistency for the variance estimators

We now turn to the investigation of consistency for the variance estimators. We shall see that under fairly weak conditions we are able to prove not only consistency (the weak form of convergence in probability), but also strong consistency (convergence almost surely) for these estimators. As mentioned, we will be working with normalized quantities. It is not very interesting to work with the variance estimators of the form $(1/n)\widehat{V}$ directly, since these will almost sure converge to 0. Thus, even if the variance estimators included in the scheme are on the form $(1/n)\widehat{V}$, the we shall work with the base estimators $\widehat{V}$. To show that all the variance estimators are consistent for their estimands, we will assume a couple of regularity conditions for the situations we are working within. The assumption goes as follows:

Assumption 3.4.2. *Let $Y_1, \ldots, Y_n$ be iid variables from a distribution with cdf G. Let μ be a one dimensional focus parameter, and θ the p-dimensional parameter vector of the parametric family of distributions with cdf F_θ and least false parameter θ_0. For this situation assume the following:*

(i) There exist a neighborhood $\mathcal{N}$ of θ_0 where $U(y;\theta)$ and $I(y;\theta)$ are continuous in θ for all $y \in \Omega$, the sample space of Y_i.

(ii) $\mathrm{IF}_\mu(y;H)$ *is continuous and bounded by an integrable function $K_0(y)$ for every cdf H.*

(iii) Letting $\kappa_1(y,\theta) = U(y;\theta)U(y;\theta)^t, \kappa_2(y,\theta) = I(y;\theta)$ and $\kappa_3(y,\theta) = U(y;\theta)\mathrm{IF}_\mu(y;H)$, there exists an integrable function $K_1(y) \geq 0$ such that

$$E_G\left[\|K_1(Y_i)\|\right] < \infty \text{ and } \|\kappa_I(y;\theta)\| \leq K_1(y) \text{ for } i = 1,2,3, \text{ and all } y,$$

where the condition for κ_3 holds for any cdf H.

(iv) The functional $\mu(H)$ is Gâteaux differentiable at both G and $\widehat{G}_n$.

(v) $\sup_{|y|\leq c}\left|\mathrm{IF}_\mu(y;\widehat{G}_n) - \mathrm{IF}_\mu(y;G)\right| = o_p(1)$ *for any $c > 0$.*

(vi) There exists a $c_0 > 0$ and a function $K_2(y) \geq 0$ such that

$$E_G[|K_2(Y_i)|] < \infty \qquad Pr\left\{\left(\mathrm{IF}_\mu(y;\widehat{G}_n)\right)^2 \leq K_2(y) \text{ for all } |y| \geq c_0\right\} \to 1.$$

(vii) $\sup_{|y|\leq c}\left(\|U(y;\theta)\|\left|\mathrm{IF}_\mu(y;\widehat{G}_n) - \mathrm{IF}_\mu(y;G)\right|\right) = o_p(1)$ *for any $c > 0$ and any $\theta \in \mathcal{N}$.*

(viii) There exists a $d_0 > 0$ and a function $K_3(y) \geq 0$ such that

$$E_G[|K_3(Y_i)|] < \infty \qquad Pr\left\{\left\|U(y;\theta)\mathrm{IF}_\mu(y;\widehat{G}_n)\right\| \leq K_3(y) \text{ for all } |y| \geq d_0\right\} \to 1.$$

(ix) $\mu_F(\theta)$ is continuously differentiable for all $\theta \in \mathcal{N}$.

To prove consistency of the variance estimators, we will have to deal with uniform convergence of functions that are means of variables depending on n. This is fortunately not an entirely new approach. Therefore, parts of the derivations will be rephrasing from textbooks on mathematical statistics. The main argument of the parametric part of the derivations is solved by applying what we call Le Cam's uniform convergence theorem (B.2.11). The standard version (from our source) is however applicable only in the situation where $p = 1$. We therefore expand this theorems to be able to handle not only scalars, but also vectors or even more generally matrices.

Corollary 3.4.3. *When the conditions of lemma B.2.11 holds for $\theta \in \Theta$ and every element $S_{(i,j)}(y,\theta)$ of a $r \times s$ dimensional matrix function $\mathcal{S}(y,\theta) = \left[S_{(i,j)}(y,\theta)\right]_{i=1,\ldots,r,\ j=1,\ldots,s}$, where $\xi_{\mathcal{S}}(\theta) = E_G[\mathcal{S}(Y_i,\theta)]$ with elements $\xi_{S,(i,j)}(\theta)$. Let now $\|\cdot\|$ denote the Frobenius norm $\|A\| = \sqrt{\sum_{i=1}^r \sum_{j=1}^s a_{i,j}^2}$ where $a_{i,j}$ are the elements of the matrix A. Then*

$$Pr\left\{\lim_{n\to\infty}\sup_{\theta\in\Theta}\left\|\frac{1}{n}\sum_{i=1}^n \mathcal{S}(Y_i,\theta) - \xi_{\mathcal{S}}(\theta)\right\| = 0\right\} = 1,$$

i.e.

$$\sup_{\theta\in\Theta}\left\|\frac{1}{n}\sum_{i=1}^{n}\mathcal{S}(Y_i,\theta)-\xi_{\mathcal{S}}(\theta)\right\|\overset{a.s.}{\to}0.$$

Before we go on to prove this we note that the Frobenius norm is simply a generalization of the Euclidean norm for vectors which again generalizes the absolute value for scalars.

Proof. Since lemma B.2.11 holds for every element of $\mathcal{S}(y,\theta)$, we have for each element (i,j) that

$$Pr\left\{\lim_{n\to\infty}\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n}S_{(i,j)}(Y_i,\theta)-\xi_{S,(i,j)}(\theta)\right|=0\right\}=1.$$

This implies that

$$Pr\left\{\lim_{n\to\infty}\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n}S_{(i,j)}(Y_i,\theta)-\xi_{S,(i,j)}(\theta)\right|=0,\forall i=1,\ldots,r,\ j=1,\ldots,s\right\}=1,$$

since for two events A and B happening with probability 1, we have

$$Pr\{A\cap B\}=Pr\{A\}+Pr\{B\}-Pr\{A\cup B\}=1+1-1=1,$$

which easily can be generalized to more than two events. Let now $(1/n)\sum_{i=1}^{n}\mathcal{S}(Y_i,\theta)=\bar{\mathcal{S}}$, with elements $\bar{\mathcal{S}}_{(i,j)}$, we get from the triangle inequality that

$$\begin{aligned}\lim_{n\to\infty}\sup_{\theta\in\Theta}\left\|\bar{\mathcal{S}}-\xi_{\mathcal{S}}(\theta)\right\| &= \lim_{n\to\infty}\sup_{\theta\in\Theta}\sqrt{\sum_{i=1}^{r}\sum_{j=1}^{s}\left(\bar{\mathcal{S}}_{(i,j)}-\xi_{\mathcal{S},(i,j)}\right)^2}\\ &\leq \lim_{n\to\infty}\sup_{\theta\in\Theta}\sum_{i=1}^{r}\sum_{j=1}^{s}\left|\bar{\mathcal{S}}_{(i,j)}-\xi_{\mathcal{S},(i,j)}\right|\\ &= \sum_{i=1}^{r}\sum_{j=1}^{s}\lim_{n\to\infty}\sup_{\theta\in\Theta}\left|\bar{\mathcal{S}}_{(i,j)}-\xi_{\mathcal{S},(i,j)}\right|.\end{aligned}$$

Now, noting that the latter is equal to zero if each of the elements inside the sum equals zero, and that this implies that the former equals 0, we get that

$$\begin{aligned}&Pr\left\{\lim_{n\to\infty}\sup_{\theta\in\Theta}\left\|\bar{\mathcal{S}}-\xi_{\mathcal{S}}(\theta)\right\|=0\right\}\\ &\geq Pr\left\{\lim_{n\to\infty}\sup_{\theta\in\Theta}\left|\bar{\mathcal{S}}_{(i,j)}-\xi_{S,(i,j)}(\theta)\right|=0,\forall i=1,\ldots,r,\ j=1,\ldots,s\right\}=1,\end{aligned}$$

which completes the proof. □

To take care of consistency due to nonparametrics, the following theorem is useful.

Theorem 3.4.4. (Almost sure convergence for means of influence functions, rewritten from Shao (2003, theorem 5.15))
Let $Y_1,\ldots Y_n$ be iid random variables from a distribution with cdf G, and let μ be a univariate

functional, which is Gâteaux differentiable at G and $\widehat{G}_n$, and let its influence function evaluated in the point y for the cdf H be given by $\mathrm{IF}_\mu(y;H)$. Suppose that

$$\sup_{|y|\leq c}\left|\mathrm{IF}_\mu(y;\widehat{G}_n)-\mathrm{IF}_\mu(y;G)\right|=o_p(1),$$

for any $c>0$ and that there exists a constant c_0 and a function $h(y)\geq 0$ such that

$$E_G[h(Y_i)]<\infty \qquad \textit{and} \qquad Pr\left\{\left|\mathrm{IF}_\mu(y;\widehat{G}_n)^2\right|\leq h(y) \textit{ for all } |y|\geq c_0\right\}\to 1.$$

Then,

$$\frac{1}{n}\sum_{i=1}^n \mathrm{IF}_\mu(Y_i;\widehat{G}_n)^2 \overset{a.s.}{\to} \int \mathrm{IF}_\mu(y;G)^2 dG(y).$$

Proof. The proof is given in Shao (2003, theorem 5.15) and consists of a rather easy part using the strong law of large numbers (theorem B.2.1) and a more complicated part where the means over $\mathrm{IF}_\mu(Y_i;\widehat{G}_n)^2-\mathrm{IF}_\mu(Y_i;G)^2$ are treated by differing on the situations where the absolute value of Y_i is greater and smaller than some constant c. □

The following lemma provides strong consistency for the normalized variance estimators included in the proposed FIC formulae of this chapter.

Lemma 3.4.5. *In a situation where the focus parameter and the set of parametric functions satisfies assumption 3.4.2, all the variance estimators $\widehat{V}_{\mathrm{np}},\widehat{V}_{\mathrm{pm}}$ and $\widehat{V}_{\mathrm{pm,np}}$, and thus also $\widehat{V}_{\mathrm{b}}$, are strongly consistent for their respective estimands.*

Proof. We start out by noting that all these estimators consist of the following five base estimators: $\widehat{J},\widehat{K},\widehat{\nu},\widehat{Q}$ and $\frac{\partial\mu_F}{\partial\theta}\big|_{\widehat{\theta}_n}$. Showing that each of these converges almost surely to their respective estimands will complete the proof by Slutsky's theorem (B.2.6) since the variance estimators simply consists of sums and matrix products of these five estimators. It is also seen by the continuous mapping theorem (B.2.9) that the last base estimator is consistent since it is by assumption continuous in θ in a neighborhood of θ_0. The main part of the proof therefore consists of showing consistence for the four first base estimators, which all are means of functions of iid variables varying with n.

We will start out by considering the estimators $\widehat{J}$ and $\widehat{K}$, which have summands depending only on $\widehat{\theta}_n$, not $\widehat{G}_n$. By using the notation from assumption 3.4.2, and also letting $\xi_j(\theta)=E_G[\kappa_j(Y_i,\theta)]$, we have for $j=1,2$ that

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^n \kappa_j(Y_i,\widehat{\theta}_n)-\xi_j(\theta_0)\right| \\
&\leq \left\|\frac{1}{n}\sum_{i=1}^n \kappa_j(Y_i,\widehat{\theta}_n)-\xi_j(\widehat{\theta}_n)\right\|+\left\|\xi_j(\widehat{\theta}_n)-\xi_j(\theta_0)\right\| \\
&\overset{*}{\leq} \sup_{\theta\in\mathcal{N}}\left\|\frac{1}{n}\sum_{i=1}^n \kappa_j(Y_i,\theta)-\xi_j(\theta)\right\|+\left\|\xi_j(\widehat{\theta}_n)-\xi_j(\theta_0)\right\|. \qquad (3.20)
\end{aligned}$$

where the $*$ over the last inequality sign denotes that the inequality holds provided that n is large enough for $\widehat{\theta}_n$ to be in $\mathcal{N}$. Since we have already seen that $\widehat{\theta}_n\overset{a.s.}{\to}\theta_0$, there will with

probability 1 exist a positive number n_0 such that $\widehat{\theta}_n \in \mathcal{N}$ for all $n > n_0$. Thus, the inequality holds with probability 1 as n increases. We can therefore without any validity loss assume n this big, and it suffices to show almost sure convergence for the two terms in expression (3.20). Almost sure convergence of the first term of expression (3.20) follows directly from corollary 3.4.3 by letting $\Theta = \mathcal{N}$. To deal with the second term we apply Lebesgue dominated convergence theorem (B.2.13) using G as the measure. The result follows by applying the triangle inequality to the result of the theorem since $\kappa_j(y, \theta)$ is bounded by an function $K_1(y)$ integrable with respect to G as a measure. Since both terms of expression (3.20) converges almost surely to zero, we have proven strong consistency for both $\widehat{J}$ and $\widehat{K}$ as estimators of J and K. Furthermore, matrix inversion is a continuous operation and therefore is also $\widehat{J}^{-1}$ strongly consistent by applying theorem B.2.9 once again.

Next we deal with the consistency of $\widehat{\nu}$. This is simple as theorem 3.4.4 shows almost sure convergence for $\widehat{\nu}$ against ν when inserting μ for the functional T. The conditions of this theorem are found in assumption 3.4.2. Thus, strong consistency of $\widehat{\nu}$ as an estimator of ν follows.

Finally we will deal with consistency of $\widehat{Q}$. Since this estimator has a summand depending on both $\widehat{\theta}_n$ and $\widehat{G}_n$, special treatment of this estimator is called for. Consistency for this estimator will be proven by wisely splitting the sum up in several terms and treating each term in different ways similar to purely nonparametric and parametric estimators. For our convenience, we introduce the notation $\widehat{Q}(\theta, H) = \frac{1}{n}\sum_{i=1}^{n} U(Y_i; \theta)\mathrm{IF}_\mu(Y_i; H)$ and $Q(\theta, H) = \int U(y; \theta)\mathrm{IF}_\mu(Y_i; H)dG(y)$. Using this notation, we have that

$$\begin{aligned}
&\left|\widehat{Q}(\widehat{\theta}_n, \widehat{G}_n) - Q(\theta_0, G)\right| \\
&\leq \left\|\widehat{Q}(\widehat{\theta}_n, \widehat{G}_n) - Q(\widehat{\theta}_n, \widehat{G}_n)\right\| + \left\|Q(\widehat{\theta}_n, \widehat{G}_n) - Q(\theta_0, G)\right\| \\
&\overset{*}{\leq} \sup_{\theta \in \mathcal{N}} \left\|\widehat{Q}(\theta, \widehat{G}_n) - Q(\theta, \widehat{G}_n)\right\| + \left\|Q(\widehat{\theta}_n, \widehat{G}_n) - Q(\theta_0, G)\right\|, \qquad (3.21)
\end{aligned}$$

where again $*$ denotes that the inequality holds with probability one when n is sufficiently large. The first term of expression (3.21) is handled by applying a version of theorem 3.4.4 to $\widehat{Q}(\theta, \widehat{G}_n)$. If we are able to show that for any $\theta \in \mathcal{N}$, $\left\|\widehat{Q}(\theta, \widehat{G}_n) - Q(\theta, \widehat{G}_n)\right\| \overset{a.s.}{\to} 0$, we have shown this also for the supremum. This follows since $\mathcal{N}$ is compact and $\left\|\widehat{Q}(\theta, \widehat{G}_n) - Q(\theta, \widehat{G}_n)\right\|$ is a real valued continuous function in θ. As a result, the function's supremum is attained by some $\theta \in \mathcal{N}$, by theorem B.2.15. We can therefore focus only on a general $\theta \in \mathcal{N}$ in the rest of the proof. We see next that condition (*vii*) and (*viii*) of assumption 3.4.2 are exactly the same conditions needed for theorem 3.4.4, if replacing one of the $\mathrm{IF}_\mu(y; \widehat{G}_n)$'s by $U(y; \theta)$. Investigation of the proof of this theorem indicates that this causes no additional problems and that the result of the theorem holds also after this replacement. As a result

$$\left\|\widehat{Q}(\theta, \widehat{G}_n) - Q(\theta, \widehat{G}_n)\right\| \overset{a.s.}{\to} 0,$$

for every $\theta \in \mathcal{N}$, which shows that the first term of equation (3.21) converges almost surely to zero.

The second term of equation (3.21) is handled by once again apply Lebesgue dominated convergence theorem (B.2.13) with G as the measure. Almost sure convergence to zero follows

by applying the triangle inequality to the result of the theorem since both $U(y;\theta)\text{IF}_\mu(y;H)$ is by assumption bounded by a function integrable with respect to G.

Since we have now shown almost sure convergence towards zero for both terms of equation (3.21), we have shown almost sure convergence to zero for $\left|\widehat{Q}(\widehat{\theta}_n, \widehat{G}_n) - Q(\theta_0, G)\right|$, and thus that $\widehat{Q}(\widehat{\theta}_n, \widehat{G}_n)$ is a strongly consistent estimator for $Q(\theta_0, G)$. By the argument in the beginning of this proof, the proof is now completed as all five base estimators are proven to be strongly consistent for their respective estimands. □

Remark 1. *The consistency result of $\widehat{Q}$ may also be handled in a more convenient way. The problematic momentum of this quantity is that is a sum depending both of $\widehat{G}_n$ and $\widehat{\theta}_n$. However, $\widehat{\theta}_n$ may be written as $\widehat{\theta}_n = T(\widehat{G}_n)$ for a certain functional T, where $\theta_0 = T(G)$. Such a functional may for H be defined as a minimizer of the Kullback–Leibler divergence between h and f_θ, where h and f_θ are the densities or pmfs of H and F_θ. Using such a representation, we may write the covariance estimator $\widehat{Q} = (1/n)\sum_{i=1}^n h(Y_i, \widehat{G}_n)$, which should work out under reasonable assumption by minor modifications of theorem 3.4.4.*

3.4.3 Consistency for the squared bias estimators

The estimators of the squared bias possess, as the variance estimators, the property of strong consistency at least in an asymptotic sense and under some additional regularity conditions. Assume the following conditions hold:

Assumption 3.4.6. *(i) The function $\mu_F(\theta)$ is continuous in the parameter θ for θ satisfying $\|\theta - \theta_0\| < \epsilon$ for some $\epsilon > 0$.*

(ii) The focus parameter as a functional $\mu(H)$ is continuous in the cdf H for H satisfying $\|H - G\|_* < \epsilon$ for some norm $\|\dot{\|}_*$ and some $\epsilon > 0$.

The true bias of the focus parameter estimators depends on n. Therefore, consistency will be dealt with in a first order asymptotic way. We define bias* quantities which we will show that the bias estimators are consistent for. This will be done in each case by rewriting the quantity inside the expectation of the bias definition and omitting the $o_p(n^{-1/2})$ terms. For large n, we have approximately that $\widehat{\mu}_{\text{pm}} - \mu_{\text{true}} \overset{d.}{=} \Lambda_{\text{pm}}/\sqrt{n} + b + o_p(1/\sqrt{n})$. Let now $\text{bias}^*_{\text{pm}} = E_G[\Lambda_{\text{pm}}/n + b] = b$ represents bias* in the parametric case. In the nonparametric case, we have analogously that approximately $\widehat{\mu}_{\text{np}} - \mu_{\text{true}} \overset{d.}{=} \Lambda_{\text{np}}/\sqrt{n} + o_p(1/\sqrt{n})$, and we therefore let $\text{bias}^*_{\text{np}} = E_G[\Lambda_{\text{np}}/n] = 0$ represent bias* in the nonparametric case. The squared versions of the bias terms, b^2 and 0, are consequently the quantities of interest. Since they are both independent of the sample size n, we can use the standard notion of strongly consistency towards these estimands. Starting with the estimate of the squared bias in the parametric case, we would like to prove that

$$\widehat{b}^2 - \frac{1}{n}\widehat{V}_b \overset{a.s.}{\to} b^2. \tag{3.22}$$

We have already seen that from lemma 3.4.1 that $\widehat{\mu}_{\text{pm}} \overset{P}{\to} \mu_{0,\text{pm}}$ (and $\widehat{\mu}_{\text{pm}} \overset{a.s.}{\to} \mu_{0,\text{pm}}$ under stronger conditions), in addition to $\widehat{\mu}_{\text{np}} \overset{a.s.}{\to} \mu_{\text{true}}$. Thus, using the continuous mapping theorem (B.2.9) once more, we get $\widehat{b}^2 = (\widehat{\mu}_{\text{pm}} - \widehat{\mu}_{\text{np}})^2 \overset{a.s.}{\to} (\mu_{0,\text{pm}} - \mu_{\text{true}})^2 = b^2$. Finally, since $\widehat{V}_b \overset{a.s.}{\to} V_b$ as shown in the previous section, $\frac{1}{n}\widehat{V}_b \overset{a.s.}{\to} 0$ and equation (3.22) follows. In the nonparametric case, the consistency is obvious since both the estimator and the estimand are zero for all n.

3.4.4 Choosing the squared bias estimator

As seen above the estimators for the squared bias have the property of being strongly consistent in an asymptotic sense. In the parametric case, we saw that both the lazy man's estimator for the parametric squared bias ($\widehat{b^2}$) and the more careful estimator that adjusts for overshooting (as in equation (3.22)), possesses this property. So in this sense, they are both "equally good". Also, the argument for using the more involved variance adjusting estimator used in section 3.2.2 was, although reasonable and intuitive, of the somewhat heuristic kind. We are therefore now going to deal with the subject in more detail. We argue that the variance adjusting estimator is better by investigating the asymptotic bias of the estimators for the squared bias. It is usual to define asymptotic bias for an estimator $\widehat{\eta}_n$ and an estimand η as $E_G[Z]/a_n$ whenever $a_n(\widehat{\eta}_n - \eta) \xrightarrow{L} Z$ for some increasing sequence a_n. This is the definition we will use as well. Obviously the estimator for the nonparametric bias estimator has this property, since the limiting squared bias is 0. Since the estimator in the parametric case consists of terms with different convergence rates, this is not a straightforward task. However, for the case where $b = 0$ we are able to derive a precise formula. From the last relation of corollary 3.1.3 we get that when $b = 0$,

$$\sqrt{n}\widehat{b} \xrightarrow{L} \Lambda_b = \sqrt{V_b}Z_0 \stackrel{d.}{=} \sqrt{V_b}N(0,1).$$

Now, using the continuous mapping theorem for convergence in law (B.2.9), we get that

$$n\widehat{b^2} \xrightarrow{L} \Lambda_{b1}^2 \stackrel{d.}{=} V_b\chi_1^2.$$

Furthermore

$$n\widehat{\text{bias}^2}_{\text{pm}} = n\left(\widehat{b^2} - \frac{1}{n}\widehat{V}_b\right) \xrightarrow{L} \Lambda_{b2}^2 \stackrel{d.}{=} V_b\chi_1^2 - V_b.$$

As a result, the asymptotic bias for the straightforward lazy man's estimator is in this case $E_G[\Lambda_{b1}^2]/n = (V_b)/n$, and the more careful variance adjusting estimator has an asymptotic bias of $E_G[\Lambda_{b2}^2]/n = (V_b - V_b)/n = 0$. In sense of asymptotic bias of the first order asymptotic approximation of the estimand, the adjusting estimator is considered a better estimator for each n when $b = 0$.

Although this result holds for the rare case when $b = 0$, there is no guarantee that it will hold in general or even for b close to zero. As mentioned the result does not generalize easily to other situations in this framework. To carry out "something" for the other situations, consider a framework where the bias reduces with increased sample size n. Especially, working with a parametric bias of the form $b_n = \Delta/\sqrt{n}$ for some $\Delta \in \mathbb{R}$ independent of n, will turn out to be quite fruitful. In particular, this situation appears in the framework where the parametric model is so-called locally misspecified. The framework we shall be working under has a distribution with density (or pmf) given by

$$g_n(y) = f(y;\theta_0) + \frac{r(y)}{\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right),$$

where $r(y) : \mathbb{R} \to \mathbb{R}$ is assumed to be a function independent on the sample size n, not necessarily continuous, but with the property that $\int r(y)\,\mathrm{d}\upsilon(y) = 0$. As usual $f(y;\theta)$ is the density (or pmf) of the parametric distribution included in the FIC scheme. This framework is investigated further in appendix A, where the limiting distribution of $\sqrt{n}\widehat{b}$ also is determined.

In such frameworks it is natural to define asymptotic bias for an estimator $\widehat{\eta}_n$ and an estimand η_0 as $\beta(\widehat{\eta}_n) = E[Z]/a_n - \eta_0$, whenever $a_n\widehat{\eta}_n \xrightarrow{L} Z$ for some increasing sequence a_n.

Under the mild regularity conditions given in appendix A, we have that

$$\sqrt{n}\widehat{b} \xrightarrow{L} N(\Delta, V_b^*), \tag{3.23}$$

where $V_b^* = V_{\text{pm}}^* + V_{\text{np}}^* - 2V_{\text{pm,np}}^*$. The derivation and explicit expressions for the quantities involved are given in corollary A.0.3. Writing the right side of equation (3.23) as $\sqrt{V_b}(Z_0 + \Delta/\sqrt{V_b})$, where $Z_0 \sim N(0,1)$, it follows from the continuous mapping theorem (B.2.9) that

$$n\widehat{b}^2 \xrightarrow{L} V_b(Z_s + \Delta/\sqrt{V_b})^2, \tag{3.24}$$

where the second factor on the right hand side can be recognized as a noncentral chi-squared distributed variable. Note also that $\widehat{V}_b \xrightarrow{P} V_b^*$ also in this framework when working under regularity conditions similar to those of assumption 3.4.2. Using relation (3.24) in addition to the consistency of $\widehat{V}_b$, we get that

$$\begin{aligned}
n\widehat{\text{bias}^2}^*_{\text{pm}} &= n\widehat{b}^2 \xrightarrow{L} V_b(Z_s + \Delta/\sqrt{V_b})^2, \\
n\widehat{\text{bias}^2}_{\text{pm}} &= n\left(\widehat{b}^2 - \frac{1}{n}\widehat{V}_b\right) \xrightarrow{L} V_b(Z_s + \Delta/\sqrt{V_b})^2 - V_b^*,
\end{aligned}$$

and thus also

$$\begin{aligned}
\beta(\widehat{\text{bias}^2}^*_{\text{pm}}) &= E[V_b^*(Z_s + \Delta/\sqrt{V_b^*})^2]/n - \Delta^2/n = V_b^*(1 + \Delta^2/V_b^*)/n - \Delta^2/n = V_b^*/n, \\
\beta(\widehat{\text{bias}^2}_{\text{pm}}) &= E[V_b^*(Z_s + \Delta/\sqrt{V_b^*})^2 - V_b^*]/n - \Delta^2/n = V_b^*(1 + \Delta^2/V_b^* - 1)/n - \Delta^2/n = 0.
\end{aligned}$$

This means that also in a local asymptotics point of view, the estimator adjusting for the overshooting of the squared bias performs better for each n. Also here "better" is in the sense of being asymptotically unbiased as an estimator of the first order asymptotic approximation of the squared bias. The estimator which is not adjusting for the overshooting will on the other hand have a nonzero bias whenever $\Delta \neq 0$, in which case we are back to $b = 0$ where we already have proved the same result in the standard framework.

3.5 Asymptotic behavior of FIC

As made clear in the sections above, the model (or estimator) with the smallest FIC score is according to the criterion of this chapter considered the best for estimating the focus parameter μ. It is certainly of interest to investigate the behavior of the presented FIC scheme. In this section we shall investigate this behavior by the use of asymptotics. The behavior will be investigated under different assumptions about the true underlying distribution.

Firstly, assume without loss of generality that we have only one parametric model pm. Considering the unimproved criterion, the nonparametric model is winning in the cases where:

$$\begin{aligned}
\text{FIC}(\widehat{\mu}_{\text{np}}) < \text{FIC}(\widehat{\mu}_{\text{pm}}) &\Leftrightarrow \frac{1}{n}\widehat{V}_{\text{np}} \leq (\widehat{\mu}_{\text{pm}} - \widehat{\mu}_{\text{np}})^2 - \frac{1}{n}\widehat{V}_{\text{np}} + 2\widehat{V}_{\text{pm,np}}, \\
&\Leftrightarrow (\sqrt{n}\widehat{b})^2 = (\sqrt{n}(\widehat{\mu}_{\text{pm}} - \widehat{\mu}_{\text{np}}))^2 \geq 2(\widehat{V}_{\text{np}} - \widehat{V}_{\text{pm,np}}).
\end{aligned} \tag{3.25}$$

Hence, we get

$$\text{FIC}(\widehat{\mu}_{\text{np}}) < \text{FIC}(\widehat{\mu}_{\text{pm}}) \Leftrightarrow \begin{cases} \sqrt{n}|\widehat{b}| > \left(2\left(\widehat{V}_{\text{np}} - \widehat{V}_{\text{pm,np}}\right)\right)^{\frac{1}{2}} & , \widehat{V}_{\text{np}} \geq \widehat{V}_{\text{pm,np}} \\ \text{always} & , \text{otherwise.} \end{cases}$$

Thus, the nonparametric estimator is the best whenever the difference between the estimators are large compared to the variation of $\widehat{\mu}_{\text{np}}$, as well as one will always choose the nonparametric estimator when the estimated covariance between the estimators are greater than the estimated variance of the nonparametric estimator(!). For the situation with several parametric models, $\widehat{\mu}_{\text{np}}$ is chosen whenever condition (3.25) holds for all parametric models "pm". If condition (3.25) does not hold for exactly one of the parametric models, this parametric model will be chosen. If this condition does not hold for more than one of the parametric models, it is easiest to see which model that is selected by checking the FIC values of these models directly.

Returning to the situation with only one parametric model, it would certainly be interesting not only to know when the different models wins in terms of the estimators, but also the probability for this to happen in certain situations. For the rest of the section we will consider model selection between one single parametric distribution and the nonparametric model. We will in the following three subsection consider three different cases, first the situation where the parametric model is fully correct, then the situation when the parametric model is locally misspecified, and finally the situation where the parametric model is misspecified for all n.

3.5.1 Selection probability under parametric truth

Reducing the model selection scheme to the statement of inequality (3.25) motivates the link between model selection and hypothesis testing. From this inequality it is natural to think of the FIC scheme as a focused hypothesis test of the null hypothesis $H_0 : G = F_{\theta_0}$, against the two-sided alternative $H_A : G \neq F_{\theta_0}$ which rejects H_0 whenever the condition (3.25) is fulfilled. Such a test will have test level given by the probability that the condition is fulfilled under the assumption that the parametric model is fully correct. As usual, the level of a test is hard to calculate when no distributional assumptions are made, and we therefore turn to asymptotics and calculate the assumption level. The following lemma provides a new joint limiting distribution for $\widehat{\mu}_{\text{np}}$ and $\widehat{\mu}_{\text{pm}}$, in the special situation when the competing parametric model is fully correct.

Lemma 3.5.1. *Let the true distribution of the iid variables $Y_1, \ldots, Y_n$ be the parametric distribution with cdf F_{θ_0} for some θ_0 in the interior of Θ. Assume furthermore that the conditions of assumption 3.1.1 are fulfilled in addition to any of the conditions of lemma 3.3.1. In addition, assume also*

$$\frac{\partial}{\partial\theta}\epsilon(\theta)\big|_{\theta=\theta_0},$$

where $\epsilon(\theta) = \mu(F_\theta) - \mu(F_{\theta_0}) - \int \text{IF}_\mu(y; F_\theta) dF_{\theta_0}$. Then

$$\sqrt{n}\begin{pmatrix} \widehat{\mu}_{\text{np}} - \mu_{\text{true}} \\ \widehat{\mu}_{\text{pm}} - \mu_{\text{true}} \end{pmatrix} \xrightarrow{L} N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} V_{\text{np}} & V^*_{\text{pm}} \\ V^*_{\text{pm}} & V^*_{\text{pm}} \end{pmatrix}\right), \tag{3.26}$$

where

$$V^*_{\text{pm}} = \left(\frac{\partial\mu_F}{\partial\theta}\Big|_{\theta_0}\right)^t J^{-1}\left(\frac{\partial\mu_F}{\partial\theta}\Big|_{\theta_0}\right),$$

and other quantities as in lemma 3.1.2.

Proof. Firstly we note the direct consequences of the fact that true distribution is the parametric distribution with cdf on the form F_θ, for some θ in the interior of Θ. We immediately see that $G = F_{\theta_0}$ and $g(y) = f(y;\theta_0)$, where θ_0 is not only the least false parameter, but may here also be named the limiting true parameter value. Now, since assumption 3.1.1 holds, the limiting distribution of lemma 3.1.2 also holds in this case. We will use this limiting distribution as a basis and show that under the additional assumptions stated, the limiting distribution is transformed to relation (3.26). Consequently from $G = F_{\theta_0}$, it follows that also $\mu_{0,\text{pm}} = \mu_{\text{true}}$. Observe also that

$$K = E_{F_{\theta_0}}[U(Y_i;\theta_0)U(Y_i;\theta_0)^t] = E_{F_{\theta_0}}[I(Y_i;\theta_0)] = J.$$

This follows by some algebra when writing out $I(y;\theta_0)$ and $U(y;\theta_0)$ in terms of $f(y;\theta_0)$ and its derivatives, cancelling terms and interchanging integration and derivation. We omit this proof since it may be found in most standard statistical textbooks, like Rice (2007). Hence, $V_{\text{pm}} = V^*_{\text{pm}}$. Thus, what remains is to prove that $V_{\text{pm,np}} = V^*_{\text{pm}}$. Since

$$V_{\text{pm,np}} = \left(\frac{\partial \mu_F}{\partial \theta}\bigg|_{\theta_0}\right)^t J^{-1}Q,$$

it is sufficient to show that $Q = \frac{\partial \mu_F}{\partial \theta}\big|_{\theta_0}$. From lemma 3.3.1, we get that

$$\mu(F_\theta) - \mu(F_{\theta_0}) = \int \text{IF}_\mu(y;F_{\theta_0})\,\mathrm{d}F_\theta(y) + \epsilon_n(\theta).$$

Differentiating this expression on both sides with respect to θ and evaluating at θ_0, we get

$$\begin{aligned}\frac{\partial \mu(F_\theta)}{\partial \theta}\bigg|_{\theta=\theta_0} &= \frac{\partial}{\partial \theta}\left(\int \text{IF}_\mu(y;F_{\theta_0})\,\mathrm{d}F_\theta(y)\right)\bigg|_{\theta=\theta_0} + \frac{\partial}{\partial\theta}\epsilon_n(\theta)\bigg|_{\theta=\theta_0},\\ &= \frac{\partial}{\partial \theta}\left(\int \text{IF}_\mu(y;F_{\theta_0})\,\mathrm{d}F_\theta(y)\right)\bigg|_{\theta=\theta_0}. \qquad (3.27)\end{aligned}$$

since by assumption $\frac{\partial}{\partial\theta}\epsilon_n(\theta)\big|_{\theta=\theta_0} = 0$. However, assuming that derivation and integration can be interchanged, we get

$$\begin{aligned}Q &= E_G[U(Y_i;\theta_0)\text{IF}_\mu(Y_i;G)] = \int U(y;\theta_0)\text{IF}_\mu(y;F_{\theta_0})\,\mathrm{d}F_{\theta_0}(y)\\ &= \int \frac{\partial}{\partial\theta}\log f(y;\theta)\bigg|_{\theta_0}\text{IF}_\mu(y;F_{\theta_0})f(y;\theta_0)\,\mathrm{d}\upsilon(y)\\ &= \int \frac{\frac{\partial}{\partial\theta}f(y;\theta)\big|_{\theta_0}}{f(y;\theta_0)}\text{IF}_\mu(y;F_{\theta_0})f(y;\theta_0)\,\mathrm{d}\upsilon(y)\\ &= \int \frac{\partial}{\partial\theta}f(y;\theta)\big|_{\theta_0}\text{IF}_\mu(y;F_{\theta_0})\,\mathrm{d}\upsilon(y)\\ &= \frac{\partial}{\partial\theta}\int \text{IF}_\mu(y;F_{\theta_0})f(y;\theta)\,\mathrm{d}\upsilon(y)\bigg|_{\theta_0}\\ &= \frac{\partial}{\partial\theta}\left(\int \text{IF}_\mu(y;F_{\theta_0})\,\mathrm{d}F_\theta(y)\right)\bigg|_{\theta_0}. \qquad (3.28)\end{aligned}$$

For this expression to hold, we must show that interchanging derivation and integration is actually a valid operation in this case. To do that we apply theorem B.2.14. The first two conditions (covering differentiability and integrability of the integrand) are clearly satisfied. The third condition concerns dominance of the derivative with an integrable function. Observe that assumption 3.1.1 assures that both $E_G[||U(Y_i;\theta_0)||^2] < \infty$ and $E_G[\mathrm{IF}_\mu(Y_i;G)^2] < \infty$, and hence both $\|U(y;\theta_0)\|^2$ and $\mathrm{IF}_\mu(y;G)^2$ are integrable. Now, since

$$\|U(y;\theta_0)\mathrm{IF}_\mu(y;G)\| \leq \|U(y;\theta_0)\|^2 + \mathrm{IF}_\mu(y;G)^2,$$

we see that $\|U(y;\theta_0)\mathrm{IF}_\mu(y;G)\|$ is dominated by an integrable function, and then the interchanging is validated. Finally, result (3.28) combined with expression (3.27) gives $Q = \frac{\partial\mu(F_\theta)}{\partial\theta}\big|_{\theta=\theta_0}$, and the proof is completed. □

Assuming the conditions of the above lemma holds, we will now investigate the probability of selecting the nonparametric model. As a direct consequence of relation (3.26), we get that

$$\sqrt{n}(\widehat{b} - b) = \sqrt{n}\widehat{b} \xrightarrow{L} N\left(0, V_{\mathrm{np}} - V_{\mathrm{pm}}\right).$$

Using Slutsky's theorem (B.2.6), and the result of lemma 3.4.5, it also follows that

$$\sqrt{n}\frac{\widehat{b}}{\sqrt{\left|\widehat{V}_{\mathrm{np}} - \widehat{V}_{\mathrm{pm,np}}\right|}} \xrightarrow{L} N(0,1),$$

since $\widehat{V}_{\mathrm{pm,np}}$ is consistent for $V_{\mathrm{pm,np}}$ and hence also for V_{pm}. As $n \to \infty$ the probability of $\widehat{V}_{\mathrm{np}} > \widehat{V}_{\mathrm{pm,np}}$ tends to one, and thus $\left|\widehat{V}_{\mathrm{np}} - \widehat{V}_{\mathrm{pm,np}}\right|$ will be consistent for $\widehat{V}_{\mathrm{np}} - \widehat{V}_{\mathrm{pm}}$. Now, considering the standard case where $\widehat{V}_{\mathrm{np}} > \widehat{V}_{\mathrm{pm}}$, we get

$$\begin{aligned}
Pr\left\{\text{select np} \mid \text{pm is true}\right\} =& Pr\left\{\mathrm{FIC}(\widehat{\mu}_{\mathrm{np}}) < \mathrm{FIC}(\widehat{\mu}_{\mathrm{pm}}) \mid \text{pm is true}\right\} \\
=& Pr\left\{\frac{2}{n}(\widehat{V}_{\mathrm{np}} - \widehat{V}_{\mathrm{pm,np}}) < (\widehat{b})^2 \mid \text{pm is true}\right\} \\
=& Pr\left\{\left(\sqrt{n}\frac{\widehat{b}}{\widehat{V}_{\mathrm{np}} - \widehat{V}_{\mathrm{pm,np}}}\right)^2 \geq 2 \,\middle|\, \text{pm is true}\right\} \\
\sim& Pr\left\{Z_0^2 \geq 2\right\} = 1 - \chi_1^2(2) \approx 0.157,
\end{aligned}$$

where $Z_0 \sim N(0,1)$ is a standard normal distributed variable.

The results show that for large n the probability is about 15.7% for choosing the nonparametric model (and estimator), when the parametric model we are fitting actually is true. I.e. in about 1 out of 6 times the nonparametric model will be chosen when the parametric model is actually fully correct. In this way the outlined model selection scheme could be seen as a focused hypothesis test, testing if the data comes from the parametric model fitted or not, with a natural asymptotic level of 0.157.

Remark 2. *The above result holds for the large class of functionals that may be written as smooth functions of averages as defined in equation* (3.17)*. The key argument of lemma 3.5.1 was that*

$$\frac{\partial\mu(F_\theta)}{\partial\theta}\bigg|_{\theta=\theta_0} = \frac{\partial}{\partial\theta}\left(\int \mathrm{IF}_\mu(y;F_{\theta_0})\,\mathrm{d}F_\theta(y)\right)\bigg|_{\theta=\theta_0}.$$

For the class of smooth function of averages, written as $\mu(H) = \phi(\int S(x)\,\mathrm{d}H(x))$*, the left side of the equation may be written as*

$$\dot{\phi}\left(\int S(x)\,\mathrm{d}F_{\theta_0}(x)\right)\frac{\partial}{\partial\theta}\left(\int S(x)\,\mathrm{d}F_{\theta}(x)\right). \tag{3.29}$$

Recalling that for this family of functionals

$$\mathrm{IF}_{\mu}(y;H) = \dot{\phi}\left(\int S(x)\,\mathrm{d}H(x)\right)\left(S(y) - \int S(x)\,\mathrm{d}H(x)\right).$$

Replacing H *by* F_{θ_0} *and integration with respect to* F_θ *gives*

$$\int \mathrm{IF}_{\mu}(y;F_{\theta_0})\,\mathrm{d}F_{\theta}(y) = \dot{\phi}\left(\int S(x)\,\mathrm{d}F_{\theta_0}(x)\right)\left(\int S(y)\,\mathrm{d}F_{\theta}(y) - \int S(x)\,\mathrm{d}F_{\theta_0}(x)\right).$$

Differentiating this expression with respect to θ *gives the same as in equation* (3.29)*. The remaining follows from the lemma.*

3.5.2 Selection probability under locally parametric misspecification

The result derived in the previous section gives useful information about the scheme and motivates the use of it through hypothesis testing. However, only a portion of the truth is revealed. It says nothing about the case when the parametric model is not fully correct. It would certainly be great to also know how often different models are chosen without this quite unreasonable assumption. To consider other situations we will have to work within a different framework. Consider the local asymptotics framework used in section 3.4.4 and treated carefully in appendix A. Then the true distribution has a density or pmf on the form

$$g_n(y) = f_{\theta_0} + \frac{r(y)}{\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right). \tag{3.30}$$

Under certain regularity conditions (see corollary A.0.3), we have seen that

$$\sqrt{n}\widehat{b} \xrightarrow{L} N(\xi_b^*, V_b^*).$$

Under the assumptions of lemma 3.5.1, except that the true distribution now has the density or pmf as given in relation (3.30), we get that $\widehat{V}_{\mathrm{np}} - \widehat{V}_{\mathrm{pm,np}}$ is a consistent estimator for $V_b^* = V_{\mathrm{np}}^* + V_{\mathrm{pm}}^* - 2V_{\mathrm{pm,np}}^* = V_{\mathrm{np}}^* - V_{\mathrm{pm,np}}^*$. For convenience we now write

$$\eta = \frac{\Delta}{\sqrt{V_{\mathrm{np}}^* - V_{\mathrm{pm,np}}^*}},$$

and let

$$Z_a = Z_0 + a \sim N(a,1).$$

Using Slutsky's theorem (B.2.6) we get that

$$\frac{\sqrt{n}\widehat{b}}{\sqrt{|\widehat{V}_{\mathrm{np}} - \widehat{V}_{\mathrm{pm,np}}|}} \xrightarrow{L} N(\eta,1) \overset{d.}{=} Z_\eta$$

$$\Rightarrow\ n\frac{(\widehat{b})^2}{|\widehat{V}_{\mathrm{np}} - \widehat{V}_{\mathrm{pm,np}}|} \xrightarrow{L} (Z_\eta)^2.$$

Working under the assumption that $\widehat{V}_{\text{np}} > \widehat{V}_{\text{pm,np}}$, we then get that

$$\begin{aligned} Pr\left\{\text{select np} \mid g_n \text{ is true}, \eta\right\} &= Pr\left\{ \frac{(\sqrt{n}\widehat{b})^2}{\widehat{V}_{\text{np}} - \widehat{V}_{\text{pm,np}}} \geq 2 \middle| \; g_n \text{ is true}, \eta \right\} \\ &\sim Pr\left\{Z_\eta^2 \geq 2\right\} = 1 - \chi^2_{1,\eta^2}(2), \end{aligned}$$

where $\chi^2_{1,\eta^2}(2)$ is the cumulative distribution function of the noncentral χ^2-distribution with noncentrality parameter η^2, evaluated at the point 2. For every known distance function from the true model $r(y)$, transformed to η, the limiting probability of selecting the nonparametric model can be calculated. Figure 3.1 shows how the probability of choosing the nonparametric model reaches 1 as η increases. Since Δ depends on both μ and the function r, we see that also

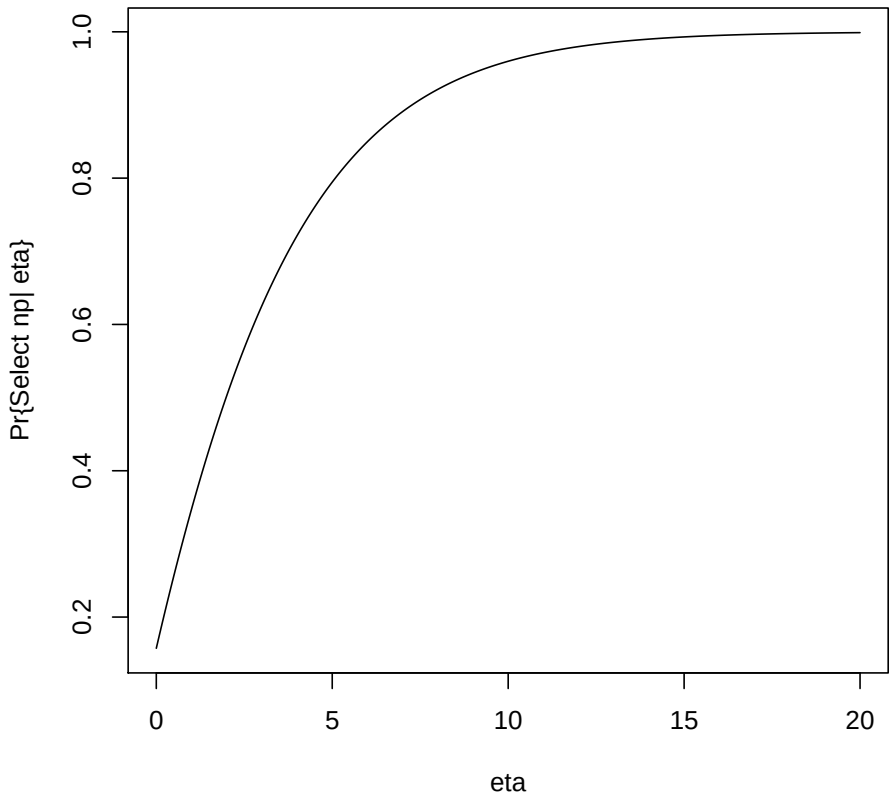


Figure 3.1: The limiting probability of choosing the estimator $\widehat{\mu}_{\text{np}}$ over $\widehat{\mu}_{\text{pm}}$ as a function of $\eta = \frac{\Delta}{\sqrt{V^*_{\text{np}} - V^*_{\text{pm,np}}}}$ when working with a locally misspecified model g_n. Note that the y-axis is on log-scale.

η and hence the probability of choosing the nonparametric model, depends on both of these as well. We also observe the natural consequence that if the variance of the nonparametric

estimator is just slightly bigger than the variance of the parametric model, η gets big and the probability of choosing the nonparametric model increases. On the other hand, if the variance of the nonparametric distribution is much larger than for the parametric model, the probability of choosing the nonparametric model is small and will reach 0.157 as the difference between the variances increases, if everything else is constant.

In a given situation one can insert the estimates $\widehat{V}_{\text{np}}, \widehat{V}_{\text{pm,np}}$ and similarly estimate the quantities involved in Δ to calculate estimates of η for different functions r. The following subsection illustrates this for the situation of local misspecification of the normal distribution.

3.5.3 Illustration: Local misspecification around the normal distribution

Consider for now the theoretical situation where model selection is performed among the normal distribution and the nonparametric distribution, where the true density is on the form given in equation (3.30). For simplicity we will assume that $\theta_0 = (0, 1)$, which corresponds to the standard normal distribution being the limiting true distribution. Furthermore we will consider a misspecification function $r(y) = f(y; \theta_0) - f_{\text{ged}}(y, 0, 1, \gamma)$, where $f_{\text{ged}}(y, \xi, \sigma, \gamma)$ denotes the density of the generalized exponential distribution with location parameter ξ, scale parameter $\sigma > 0$ and shape parameter $\gamma > 0$. This distribution has the normal distribution with parameters ξ and σ as a special cases when $\gamma = 2$. It may in this way be seen as a generalization of the normal distribution. Note furthermore that when $\gamma = 1$ the distribution corresponds to the Laplace distribution and as $\gamma \to \infty$ the distribution converges to a Uniform distribution on the interval $[\xi - \sigma, \xi + \sigma]$. For this situation we consider the focus parameter $\mu(H) = H^{-1}(0.9)$, i.e. the upper 10% quantile of the distribution (here denoted by the general cdf H). Figure 3.2 indicate the limiting selection probability for varying values of the γ parameter. The figure shows how the probability of selecting the nonparametric estimator changes as the true distribution departs from the point $\gamma = 2$ corresponding to the parametric distribution being fully correct. The smallest probability is as expected found in the point $\gamma = 2$. A value of γ smaller than 2 drastically increases the probability of selecting the nonparametric distribution, whereas a γ greater than 2 does not affect the selection probability that much. The reason for this is that when $\gamma < 2$ the tails of the distribution changes rapidly, which highly affects the variance. When $\gamma > 2$, the changes in the tails are not as significant. Similar plots may be carried out for other parametric distributions and focus parameters.

3.5.4 Selection probability under misspecified parametric models

We will finish off this section regarding selection probability by investigating which model is chosen when the true model is fixed for each n and is not exactly the parametric model that is included in the set of competing models for model selection. Also in this case we rely on asymptotics. This situation is actually the most common one, since as George Box said: "Essentially, all models are wrong, but some are useful." Since statisticians as a result always use wrong models, an assumption that a model is correct seems rather repellent. Therefore we now use a few lines to investigate what happens as n increases and the parametric model is correct. From corollary 3.1.3 we get that

$$\sqrt{n}(\widehat{b} - b) \xrightarrow{L} N(0, V_b).$$

The key argument that leads to the magical 0.157 for the case when the parametric model is fully correct is that $b = \mu_{0,\text{pm}} - \mu_{\text{true}} = 0$. Consider now the situation where $b = \mu_{0,\text{pm}} -$

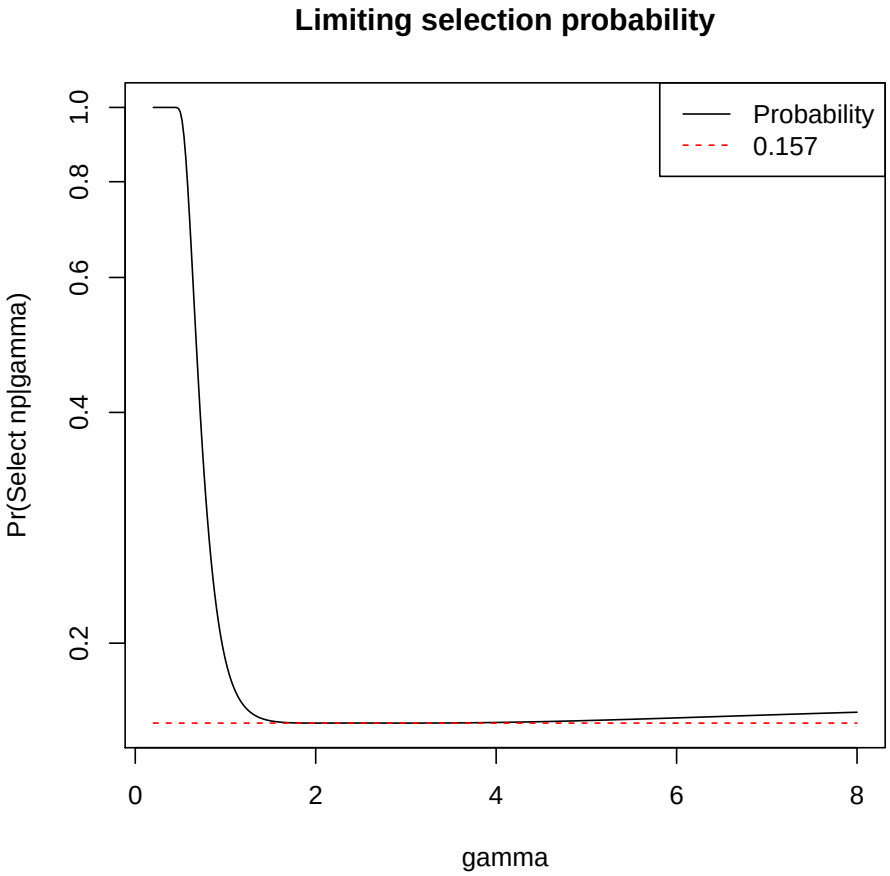


Figure 3.2: The limiting probability of choosing the estimator $\widehat{\mu}_{\text{np}}$ over $\widehat{\mu}_{\text{pm}}$ as a function of γ, the shape parameter of the generalized exponential distribution, when working with a locally misspecified model g_n.

$\mu_{\text{true}} \neq 0$. From the limiting distribution above we see that $\sqrt{n}(\widehat{b} - b)$ will grow to infinity as n increases. This is also seen from the fact that when $\mu(H)$ is continuous in the cdf H, $\widehat{b} = \mu(\widehat{G}_n) - \mu(F_{\widehat{\theta}_n}) \xrightarrow{a.s.} \mu_{\text{true}} - \mu_{0,\text{pm}} = b$. And as $b \neq 0$, $\sqrt{n}\widehat{b} \xrightarrow{a.s.} \infty$. When this is the case, the left hand side of equation (3.25) converges almost surely to ∞, whereas the right hand side still converges almost surely to $2(V_{\text{np}} - V_{\text{pm,np}})$. Therefore it follows that the probability of selecting the nonparametric estimator converges almost surely to 1 as $n \to \infty$ whenever $b \neq 0$. Even if the parametric model is almost a perfect fit for the data and $b = \epsilon$ for a very small number ϵ, the nonparametric will be chosen with probability 1 as $n \to \infty$. Directly from this it is also seen that if there are several parametric models and for all of them $b \neq 0$, then FIC tends to select the nonparametric model as n increases. This follows since with probability tending to 1, FIC_{np} will be smaller than each of the FIC_{pm}.

Consider now a situation where there is still one parametric model, that it is not the true model, but still $b = 0$. This is e.g. the case when data is $N(\xi, \sigma)$ and we only fit the model $N(\xi, 1)$. In general we then have

$$\sqrt{n}\widehat{b} \xrightarrow{L} N(0, V_b).$$

The results giving $V_{\text{b}} = V_{\text{np}} - V_{\text{pm,np}}$ does however not hold in this situation. By using arguments from the above derivations we get that as $n \to \infty$, the probability of selecting the nonparametric model whenever $b = 0$, but not necessarily $G = F_{\theta_0}$, equals

$$\begin{aligned} Pr\left\{\text{select np } | b = 0\right\} &= Pr\left\{ \frac{n\widehat{b}^2}{\widehat{V}_{\text{np}} - \widehat{V}_{\text{pm,np}}} \geq 2 \middle| b = 0 \right\}, \\ &\sim Pr\left\{ \frac{Z^2_{V_{\text{b}}}}{V_{\text{np}} - V_{\text{pm,np}}} \geq 2 \right\} = Pr\left\{ Z_0^2 \geq 2\frac{V_{\text{np}} - V_{\text{pm,np}}}{V_{\text{np}} + V_{\text{pm}} - 2V_{\text{pm,np}}} \right\}, \\ &= Pr\left\{ Z_0^2 \geq 2\left(1 - \frac{V_{\text{pm}} - V_{\text{pm,np}}}{V_{\text{np}} + V_{\text{pm}} - 2V_{\text{pm,np}}}\right) \right\} = 1 - \chi_1(2\kappa), \end{aligned} \tag{3.31}$$

where $\kappa = 1 - \frac{V_{\text{pm}} - V_{\text{pm,np}}}{V_{\text{np}} + V_{\text{pm}} - 2V_{\text{pm,np}}}$. As a result, this selection probability does not have a general answer. We do however observe that the probability of choosing the nonparametric model is larger than for the case where we also assume $G = F_{\theta_0}$, whenever $V_{\text{pm}} > V_{\text{pm,np}}$, which is most often also the case. That is also quite natural, since although the parametric model will tend to give the exact answer it is still not correct.

3.6 Performance

In this section we will discuss and investigate the performance of the main FIC schemes for iid data. We will especially compare FIC with the most commonly used information criteria AIC and BIC. This comparison will be done partly theoretical and partly via simulations for a few selected situations.

3.6.1 Limiting performance under misspecified parametric models

As shown in section 3.5 the nonparametric estimator will be chosen with probability 1 as $n \to \infty$ whenever $b = \mu_{0,\text{pm}} - \mu_{\text{true}} \neq 0$ for all parametric models. This is a very powerful property, since it ensures that the best model (the nonparametric model) will be chosen eventually as n increases. This is the case since the parametric estimator will tend to $\mu_{0,\text{pm}}$, and not the value

μ_{true} we are aiming for, which the nonparametric estimator will tend to. Thus, when $b \neq 0$ no matter what the true distribution of the data is, the nonparametric estimator will be the best as long as n gets large enough. This is clearly the case no matter how many parametric models we try to fit, as long as none of them has the property that $b = 0$ exactly.

The real power of this property come into play when comparing the proposed FIC scheme with other model selectors only dealing with parametric models. Claeskens and Hjort (2008, Chapter 4) summarizes limiting selection results for AIC and BIC. Both of these have the weak consistency property that when exactly one of the competing models minimizes the Kullback–Leibler divergence, this model will be chosen with probability 1 as $n \to \infty$. Strong consistency which is defined as the property of selecting the model with the smallest number of parameters if there are more models with this minimizing property, is however only possessed by BIC. AIC on the other hand has a bounded risk function, a property BIC does not possess. No matter what, when the correct model is not included in the set of competing models, neither can it be selected – no matter how the information criterion behaves. Thus, when the parametric models included in the set of competing models neither possess the property $b = 0$, the estimator based on the winning model will tend to a value different from μ_{true}. Therefore, FIC scheme tending to choose the best model as n increases are in this sense robust against deviation from $b = 0$, whereas information criteria including only parametric models whenever is not. In other words FIC will tend to select better models in this situation.

3.6.2 A simulation study of FIC performance

In the previous section we explained the great property that the FIC scheme for iid data will always tend to choose the correct model when $b \neq 0$ as $n \to \infty$, a property that few other model selection schemes possess. In section 3.5.1 we have also seen that when $b = 0$ and the single parametric model fitted is also true, the nonparametric model is still chosen roughly 1 out of 6 times. These properties are certainly interesting, but in practical situations the sample sizes are finite, and the results no longer apply in general.

In this section we use simulations to study how the performance of the FIC scheme is compared to AIC and BIC for a few situations where the sample size is finite. To not favor FIC, we will let the true distribution be among the parametric model fitted. Especially we will simulate data from the distribution Weib(a_0, b_0), where a_0 is the shape parameter and b_0 is the rate parameter. In this short study we let $a_0 = 1.1$ and $b_0 = 1$. The set of competing models will consist of the usual nonparametric model, the exponential model Exp(λ), and the weibull distribution Weib(a, b). We will focus on three different focus parameter in this study: The variance, the mean and the third moment of the distribution. The calculation of the FIC formulae for each situation requires some computation time since numerical approximations with high accuracy are used. Therefore we only study these situations for the three different sample sizes $n = 50, 200, 400$ and the presented results are based on only 10^4 sampled data sets. Because of the quite small number of repetitions, we stress that the results presented below are only rough approximations. Since the adjusted FIC scheme (making sure any negatively squared bias is estimated to zero instead) is most natural to apply for practical problems, we use this version all the way.

Below we give tables with the summarizing results from the simulation study. We each focus parameter and each sample size we give the quantity RMSE* for AIC, BIC and FIC. RMSE* is $\sqrt{n}$ times the mean of the absolute distance between the estimate chosen by the information criterion and the true value. The factor $\sqrt{n}$ is included to make the quantity easier to compare

for different sample sizes.

Table 3.1 below shows the simulation results when the variance is the focus parameter. As we see from the table, FIC performs best both when $n = 50$ and when $n = 200$. When $n = 400$ AIC is a better choice. All the way both AIC and FIC clearly outperforms BIC.

	RMSE*		
	$n = 50$	$n = 200$	$n = 400$
AIC	1.554	1.676	1.583
BIC	1.615	2.115	2.296
FIC	1.420	1.642	1.683

Table 3.1: Resulting AIC, BIC and FIC scores based on simulation when $\mu =$ the variance.

Table 3.2 below shows the simulation results for the mean as focus parameter. This type of model selection is maybe a little silly as the nonparametric estimator coincides with the estimator under exponential distribution. Investigating this does however have the benefit that all schemes have the opportunity to choose among the same final estimators. As we see from the table, all information criteria performs almost equally well. The main reason for this is that the estimators under the models are so similar. The results may partly be due to randomization error, but still the tendency is that FIC selects wisest. From the simulations one may in addition observe that FIC chooses the best model $2-4\%$ more often than AIC and $4-5\%$ more often than BIC. Here the best model for each simulation is defined as the model that has $\widehat{\mu}$ with smallest distance to the true value of the focus parameter. Even if the results here are quite similar, quite different models are selected. BIC most often selects the exponential model, AIC selects the exponential model for small sample sizes, but as the sample size increases it selects the Weibull model more often. FIC on the other hand does almost always select the Weibull model. In fact, when $n = 400$ all simulated data sets resulted in FIC choosing the Weibull model. This result is likely to be a consequence of the adjustment of the FIC scheme.

	RMSE*		
	$n = 50$	$n = 200$	$n = 400$
AIC	0.7007	0.7036	0.6962
BIC	0.7006	0.7038	0.6962
FIC	0.7002	0.7035	0.6961

Table 3.2: Resulting AIC, BIC and FIC scores based on simulation when $\mu =$ the mean.

Finally, table 3.3 shows the simulation results for the third moment as focus parameter. The simulations indicate that for small samples FIC performs clearly best. For $n = 400$ AIC seems to do the best job.

From these simulation it is also seen that FIC tends to underestimate the mse slightly. The FIC values and estimators of the mse are quite good, but for most situations the FIC value of the winning estimator seems to be slightly smaller than the squared distance between this value and the true value. The reason for this seems to be related to the squared bias estimator. For the nonparametric estimator it is zero and will thus always be smaller than or equal to the true squared bias for finite and infinite n. The squared bias estimator for the parametric

	RMSE*		
	$n = 50$	$n = 200$	$n = 400$
AIC	12.045	12.486	11.875
BIC	12.719	15.346	16.333
FIC	10.311	11.836	12.640

Table 3.3: Resulting AIC, BIC and FIC scores based on simulation when $\mu =$ the third moment.

μ estimator is also sometimes zero. This happens whenever the estimated variance of the bias estimator is greater than the square of the direct bias estimator. When this happens it is most often an underestimation of the true squared bias. When it does not happen it is not clear whether the estimate is greater or smaller than the true value. Thus, in the cases where we can see whether there is overestimation or underestimation, it is always underestimation. Therefore it is not a surprise that the FIC values seems to be slightly too small on average.

The reason why AIC is performing so well for large sample sizes may partly be due to the setup of the simulation study. As discussed in section 2.2, AIC tends to choose bigger models than BIC for larger sample sizes, which we also directly see from the penalizing term. Since the true model now actually is the biggest of the parametric models, AIC may be slightly favored. One may have gotten other results, if a model bigger than the true model where also included in the set of focus parameters. We do not try out or go any deeper into this as this was just a minor study to indicate how FIC perform also for finite samples.

It should also be noted that for more complex focus parameters than those studied here, FIC does not seem to perform as well as one would have hoped. The reason for this is likely found in the strength of the mse-estimator. For some more complex focus parameters the nonparametric estimator is biased for finite n. So even if the asymptotic bias is zero, the mse may be slightly underestimated in these situations. Thus, for more complex situations, second order approximations may be called for. That is however outside the scope of this thesis.

3.7 A special case

In some situations it is of interest to check or test whether a fixed distribution matches a data set well or not. Such checks are most often performed by traditional goodness of fit test. One way to think of a fixed distribution is as a parametric distribution with no parameters. We shall in this short section see that by working with the fixed distribution in such a view, one may use FIC as a focused goodness of fit test for a particular fixed distribution.

Denote by μ_0 the value of μ under some fixed distribution, and consider for simplicity the regular FIC scheme of this chapter with the nonparametric and just this fixed distribution as the set competing models. Since there are no parameters to be estimated in the parametric model, the parametric "estimator" of μ is simply a constant always equal to μ_0. Consequently this "estimator" has variance zero, and it is also immediate that the covariance with the nonparametric estimator also is zero. The parametric estimator of μ will however have a nonzero bias in general. By simply applying FIC to this situation we arrive at the following FIC formulae

$$\text{FIC}(\widehat{\mu}_{\text{np}}) = \widehat{\text{mse}}(\widehat{\mu}_{\text{np}}) = \frac{1}{n}\widehat{V}_{\text{np}},$$
$$\text{FIC}(\mu_0) = \widehat{\text{mse}}(\mu_0) = (\mu_0 - \widehat{\mu}_{\text{np}})^2 - \frac{1}{n}\widehat{V}_{\text{np}}.$$

Consequently, the nonparametric estimator is declared the winner if

$$n(\widehat{\mu}_{\text{np}} - \mu_0)^2 > 2\widehat{V}_{\text{np}}. \tag{3.32}$$

In the spirit of section 3.5, it is of interest to calculate the probability that the nonparametric model wins when $\mu_0 = \mu_{\text{true}}$. Under this working hypothesis it is easily seen from the results of section 3.5 and corollary 3.1.3 that when $\widehat{V}_{\text{np}}$ is consistent, the probability of choosing the nonparametric estimator tends to

$$Pr\left\{Z_0^2 > 2\right\} = \chi_1^2(2) \approx 0.157.$$

As a result, the focused test for the validity of the fixed distribution has asymptotic level of approximately 0.157. Note that since the only part of the fixed distribution we care about is μ_0, the test may also be seen as a plain estimate test instead of a model test. The test given in equation (3.32) is very simple and has the advantage of being theoretically motivated in addition to being a special case of the more general model selection routine of FIC.

3.8 Multivariate extension

Up until now, we have considered only univariate data. As will be made clear in this section the generalization to iid multivariate data follows without too much trouble. To do this we do however need to redefine some of the quantities used. The parametric part of the scheme does not need much modification since the assumptions and statements are quite general. For the nonparametrics some refinement is however needed. By omitting the most obvious details and straightforward generalizations for the univariate case this section will include arguments leading to the multivariate extension of the main scheme of this chapter.

3.8.1 Heuristic derivation

To derive a FIC scheme for this situation which is able to deal with multivariate data, we are first going to define a few quantities generalizing from the simpler univariate case. For this section, assume that data $Y_1, \ldots, Y_n$ are r-dimensional iid variables stemming from an r-dimensional distribution with density or pmf given by $g(y) = g(y_1, \ldots, y_r)$ and cdf $G(y) = G(y_1, \ldots, y_r)$. Furthermore, Y_i can be written in an element wise way as $Y_i = (Y_{1i}, \ldots, Y_{ri})^t$. Moreover, we assume that the focus parameter of interest is univariate and can be written as a functional of a r-dimensional cdf.

One of the crucial extensions from the univariate case is the one concerning the empirical cumulative distribution function. In the univariate case it is defined for univariate y and Y_i as $\widehat{G}_n(y) = \frac{1}{n}\sum_i^n \mathbf{1}_{\{Y_i \leq y\}}(y)$. For r-dimensional data and evaluation points $y = (y_1, \ldots, y_r)^t$, the empirical cumulative distribution function generalizes to

$$\widehat{G}_n(y_1,\ldots,y_r) = \frac{1}{n}\sum_i^n \mathbf{1}_{\{Y_{1i}\leq y_1 \cap \cdots \cap Y_{ri}\leq y_r\}}(y_1,\ldots,y_r),$$

where $\cap$ denotes the intersection of sets or logically "and". Furthermore, we define the general influence function of a functional μ at the r-dimensional cdf H in y (which is r-dimensional) as the univariate quantity

$$\mathrm{IF}_\mu(y;H) = \lim_{\epsilon\to 0}\frac{\mu(H+\epsilon(H+\delta_y))-\mu(H)}{\epsilon},$$

where the multivariate cdf of Dirac's delta measure is given by $\delta_y(x) = \mathbf{1}_{\{y_1\leq x_1\cap\cdots\cap y_r\leq x_r\}}(x)$ in the r-dimensional vector $x=(x_1,\ldots,x_r)$.

As mentioned, the crucial extension concerns the nonparametrics, not the parametrics. For the parametric part one simply needs to replace univariate Y_i with multivariate Y_i for estimators which are functions of data. In addition one must be aware that all parameters of the multidimensional parametric distribution must be placed in the single column vector of parameters θ when limiting distributions are derived. The latter yields whether the parametric distributions consist of vectors or matrices with parameters to be fitted.

Now, studying assumption 3.1.1 which is used to derive the joint limiting distribution of the parametric and nonparametric estimators, there is no assumption not making sense or behaving differently for multivariate data. We assume that this holds also in this multivariate setting. Lemma 3.1.2 which states the exact form of this joint limiting distribution uses only this assumption, the delta method, Slutsky's theorem and some algebra, which are no different when data are multivariate. Therefore, this lemma holds also for multivariate data, implying that the marginal limiting distributions of corollary 3.1.3 also hold. Using the multivariate extensions of the estimators involved in the FIC formulae presented in section 3.2 yields a FIC formulae and a FIC scheme for multivariate data on the exact same form as in the univariate situation:

$$\begin{aligned}\mathrm{FIC}(\widehat{\mu}_{\mathrm{np}}) &= \widehat{\mathrm{mse}}(\widehat{\mu}_{\mathrm{np}}) = \frac{1}{n}\widehat{V}_{\mathrm{np}},\\ \mathrm{FIC}(\widehat{\mu}_{\mathrm{pm}}) &= \widehat{\mathrm{mse}}(\widehat{\mu}_{\mathrm{pm}}) = (\widehat{\mu}_{\mathrm{pm}}-\widehat{\mu}_{\mathrm{np}})^2 - \frac{1}{n}\widehat{V}_{\mathrm{np}} + 2\frac{1}{n}\widehat{V}_{\mathrm{pm,np}}.\end{aligned}$$

Finally, all consistency and unbiasedness results also hold in this situation since none of the results are directly affected by the dimension of the cdfs and Y_is. All in all, everything that holds in the univariate case, should also hold for Y_i multivariate when extra care is taken whenever one deals directly with Y_i and the different cdfs.

3.9 Examples and illustrations

We finish off this chapter by giving a few examples and illustrations.

3.9.1 Example 1: The Norwegian population's activity level

To encourage and guide the Norwegian residents towards a higher physical activity level to gain health benefits, the Norwegian Directorate of Health has among other things recommended that adults should walk a minimum of 10 000 steps each day. On behalf of the Norwegian Directorate of Health, the Norwegian School of Sport Sciences conducted a national survey of the activity level of Norwegian adults under the name Kan1. The survey consisted of a questionnaire and an electrical equipment worn by the attendees for on average a full week. The equipment recorded with high accuracy different types of activity indicators, among them all steps taken and the intensity of the activity performed. The survey was conducted from spring 2008 and was finished off in the spring 2009. In total 3464 individuals were tested, with a good mix of men and women, young and old and with a reasonable geographical spread. For further information about the Kan1 study, see the survey report Anderssen et al. (2009).

In this example we will focus on the number of daily steps walked. Our focus parameter will be the portion of the Norwegian adult population that satisfy the recommendation of at least 10 000 steps daily. This focus parameter says something about how active the Norwegian population is. To estimate this proportion, we will use data gathered from the Kan1 survey, especially we will use the average amount of steps per day for each of the individuals participating in the study, omitting individuals 65 years and older and those that have used the equipment for less than 4 days. The reason for omitting the oldest individuals is that the official recommendation is given to adults 18–64 years old, other recommendations usually apply to the elderly. Note also that no individuals are younger than 20 years old in the survey. This may underrepresent this age group slightly, but such a restriction should not cause any trouble. Returned from the filtering of the data were 2527 individuals each with a high quality measure representing the number of daily steps taken. Note also that the data we use for the number of daily steps for each individual is an average number calculated from the total number of days the equipment was worn by the individual. Adjustments were also made for time periods where the equipment was not worn. The lowest recorded value of daily steps among the 2527 individuals is 470 steps, and the greatest number is 24070 steps.

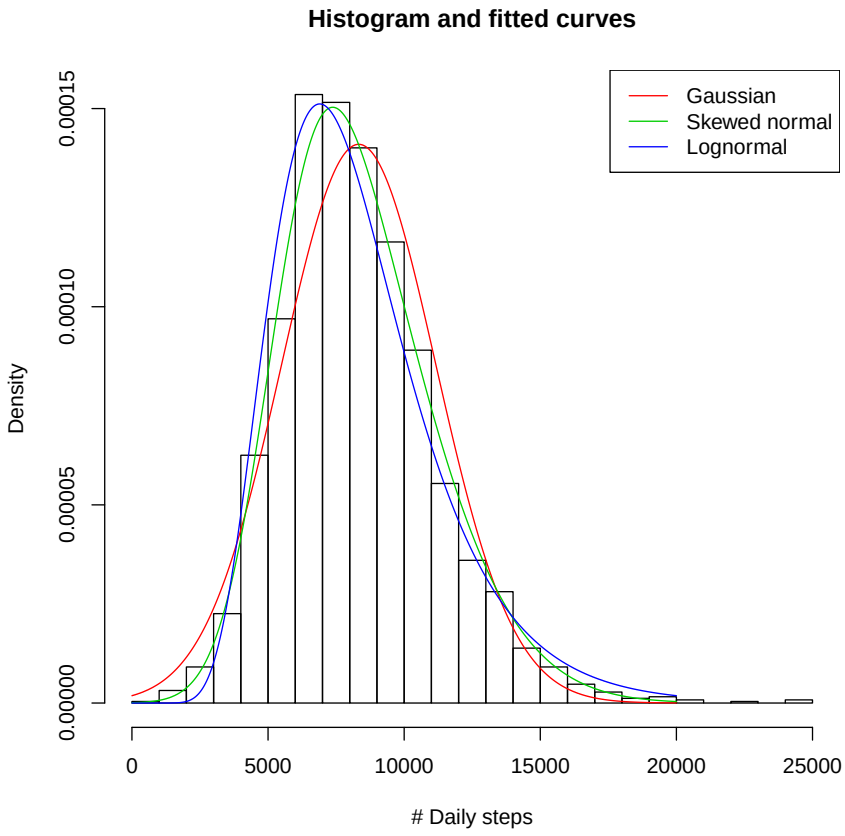
For our focus parameter there are a great number of different possibilities, where the non-parametric approach of simply counting the number of individuals that have a measured number of steps above 10 000, is highly actual. It is also natural to think of such data as normally distributed since this a population phenomenon which often turns out to have a unimodal distribution. Simply plotting a histogram of the data also verifies this heuristic "guess". The histogram does however indicate a slight skewness to the right, and we therefore propose the skewed normal distribution and the log-normal distribution. The latter is natural also since the underlying distribution should give positive probability only to the positive half of the real line. To summarize, we propose the models and estimators given in table 3.4. A histogram of

Model	Density	Cdf	μ estimator
Nonpar	Undefined	$\widehat{G}_n(y)$	$1-\widehat{G}_n(10000)$
Normal	$\phi(\frac{y-\xi}{\sigma})$	$F_{\text{norm}}(y)=\frac{1}{\sigma}\Phi(\frac{y-\xi}{\sigma})$	$1-F_{\text{norm}}(10000)$
S. normal	$f_{\text{s.n.}}(y)=\frac{2}{\sigma}\phi(\frac{y-\xi}{\sigma})\Phi(\alpha\frac{y-\xi}{\sigma})$	$F_{\text{s.n.}}(y)=\int_{-\infty}^{y} f_{\text{s.n.}}(x)dx$	$1-F_{\text{s.n.}}(10000)$
Log-normal	$\phi(\frac{\log(y)-\xi}{\sigma})/y$	$F_{\text{l.n.}}(y)=\Phi(\frac{\log(y)-\xi}{\sigma})$	$1-F_{\text{l.n.}}(10000)$

Table 3.4: Table of models and estimators fitted in example 1, where $\phi(y)$ and $\Phi(y)$ as usual represents the density and cdf of the standard normal distribution.

the data with the fitted parametric curves is also provided in figure 3.3.

Figure 3.3: Histogram of the number of daily steps for a representative group of the Norwegian adult population, with the density curves of three fitted parametric distributions.



Note that since the focus parameter can be written as a smooth function of averages, the focus parameter may be handled by the proposed FIC apparatus of this chapter. Using the main FIC scheme on this situation provides results given in table 3.5. For each model we provide the following quantities: The estimate, the dimension, an estimate $\widehat{\text{bias}}^*$, which is the root of the square bias estimate, an estimate $\widehat{\text{sd}}$ of the standard deviation, and an estimate $\widehat{\text{RMSE}}$ which is the root mean squared error. $\widehat{\text{RMSE}}$ is just the root of the FIC value. Finally the rank of the models is given.[4]

The table should be more or less self-explanatory, nevertheless we point out its main components. As seen from the last column where the rank is given, the log-normal model performs best (according to this scheme) at estimating the proportion of the population that fulfills the recommendation of 10000 daily steps. The nonparametric model is a good number two, whereas the skewed normal distribution and especially the regular normal distribution both estimate this quantity less precisely. From the RMSE column we see that the estimated error of the two best models does not differ too much and that the normal distribution is clearly not

[4]The reason for using these "rooted" quantities instead of the squared bias, variance and FIC values directly is that they now are in scale of the μ values.

	$\widehat{\mu}$	dim	$\widehat{\text{bias}}^*$	$\widehat{\text{sd}}$	$\widehat{\text{RMSE}}$	Rank
Nonpar	0.2438	Inf	0	0.0085	0.0085	2
Normal	0.2776	2	0.0332	0.0088	0.0344	4
Skewed normal	0.2562	3	0.0111	0.0073	0.0133	3
Lognormal	0.2494	2	0.0014	0.0064	0.0066	1

Table 3.5: Results of the main FIC scheme fitted to example 1 of the Norwegian population's activity level.

a good choice in this situation. Before concluding, we point out a somewhat unusual feature of this analysis. The estimated variance of the μ estimator based on the normal model is actually greater than the nonparametric estimator. Thus, using this parametric model does not give increased stability for this focus parameter. The rule of thumb is however the opposite, that parametrics gives increased stability compared to nonparametrics. This feature is an even stronger indicator that the normal distribution is a poor choice of model in this situations. The log-normal distribution is the winning model for these data, and therefore the final estimate of the proportion of the adult population that does not fulfill the national recommendations regarding daily activity level, is with four digits precision 0.2494, or 24.94%. In other words almost one out of four adults is active enough to gain health benefits over time, according to the Norwegian Directory of Health definitions.

3.9.2 Example 2: Different models for different focus parameters

The key idea of focused inference is that different models may perform best at different estimation tasks. In this example will illustrate this phenomenon by fitting a few models to a randomly generated data set and then using FIC to choose which model to trust for different focus parameters. The simulated data we use consist of 500 values, where 250 of them are generated from the standard exponential distribution (parameter $\lambda = 1$) and the other 250 values are generated from the Weibull distribution with equal shape and scale parameters ($\lambda = k = 1.3$). We fit the usual nonparametric model, the exponential model and the Weibull model to this data. We also assume interest in the statistical dispersion of the data, and will measure this dispersion with the following three focus parameters, one at a time:

(i) **Variance**: $\mu_{\text{Var}}(H) = \text{Var}_H(Y_i)$

(ii) **Mean Absolute Deviation About the Median (MADAM)**: $\mu_{\text{MADAM}}(H) = E_H[|Y_i - H^{-1}(1/2)|]$.

(iii) **InterQuartile Range (IQR)**: $\mu_{\text{IQR}}(H) = H^{-1}(3/4) - H^{-1}(1/4)$.

The variance may be written on the smooth functions of averages form and is thus applicable to our FIC apparatus. The other two are simple functions of functionals that are Hadamard differentiable, and should therefore also work out fine by the chain rule of Hadamard differentiability. For completeness we provide the analytical expressions of the influence functions for these focus parameters.

$$\begin{aligned}
\mathrm{IF}_{\mu_{\mathrm{Var}}}(x;H) &= (x - E_H[Y_i])^2 - \mu_{\mathrm{Var}},\\
\mathrm{IF}_{\mu_{\mathrm{MADAM}}}(x;H) &= |x - H^{-1}(1/2)| - \mu_{\mathrm{MADAM}},\\
\mathrm{IF}_{\mu_{\mathrm{IQR}}}(x;H) &= -\frac{\mathbf{1}_{\{x\le H^{-1}(3/4)\}}(x)}{4h(H^{-1}(3/4))} + \frac{\mathbf{31}_{\{x>H^{-1}(3/4)\}}(x)}{4h(H^{-1}(3/4))} -\\
&\qquad \left(-\frac{\mathbf{31}_{\{x\le H^{-1}(1/4)\}}(x)}{4h(H^{-1}(1/4))} + \frac{\mathbf{11}_{\{x>H^{-1}(1/4)\}}(x)}{4h(H^{-1}(1/4))} \right),
\end{aligned}$$

where h denotes the density of the distribution with cdf H.

It is important to note that these parameters do not measure the exact same feature of the data, and the results of the three are therefore hard to compare. They do however all measure how much dispersion there is in the data, even if the scales are not directly comparable.

We now turn to the results of FIC analysis of these data. The tables 3.6, 3.7 and 3.8 give the results from the FIC analysis of the three different focus parameters. In addition figure 3.4 provides a histogram of the data with curves from the fitted parametric distributions. Note that we have used the adjusted version of the main scheme. The reason for this is that the estimator of MADAM based on the Weibull distribution lead to a negatively estimated squared bias. The earlier motivated convention of setting this estimate to zero therefore seemed natural here.

	μ	dim	$\widehat{\text{bias}}^*$	$\widehat{\text{sd}}$	$\widehat{\text{RMSE}}$	Rank
Nonpar	1.0945	Inf	0.0000	0.1193	0.1193	1
Exp	1.3923	1	0.2807	0.1234	0.3066	3
Weibull	1.2002	2	0.0899	0.1298	0.1579	2

Table 3.6: Results of the adjusted FIC scheme of example 2 for focus parameter μ = the variance.

	$\widehat{\mu}$	dim	$\widehat{\text{bias}}^*$	$\widehat{\text{sd}}$	$\widehat{\text{RMSE}}$	Rank
Nonpar	0.7871	Inf	0.0000	0.0367	0.0367	2
Exp	0.8175	1	0.0192	0.0363	0.0410	3
Weibull	0.7800	2	0.0000	0.0363	0.0363	1

Table 3.7: Results of the adjusted FIC scheme of example 2 for focus parameter μ = MADAM.

As seen from the resulting tables, each of the models wins once. For the variance, the nonparametric model is a clear winner. For MADAM, the Weibull distribution performs marginally better than the nonparametric model, and clearly better than the exponential model. Finally, for IQR the exponential model is a quite clear winner in front of the nonparametric and Weibull models. Summing up, all models won once which illustrates that for the same data set, and even if the focus parameters measure a similar type of quantity, the winning model does certainly not have to be the same for all focus parameters.

	$\widehat{\mu}$	dim	$\widehat{\text{bias}^*}$	$\widehat{\text{sd}}$	$\widehat{\text{RMSE}}$	Rank
Nonpar	1.3806	Inf	0.0000	0.0899	0.0899	2
Exp	1.2963	1	0.0430	0.0575	0.0718	1
Weibull	1.2613	2	0.0990	0.0559	0.1137	3

Table 3.8: Results of the adjusted FIC scheme of example 2 for focus parameter $\mu = \text{IQR}$.

3.9.3 Example 3: The number of goals scored during a football match

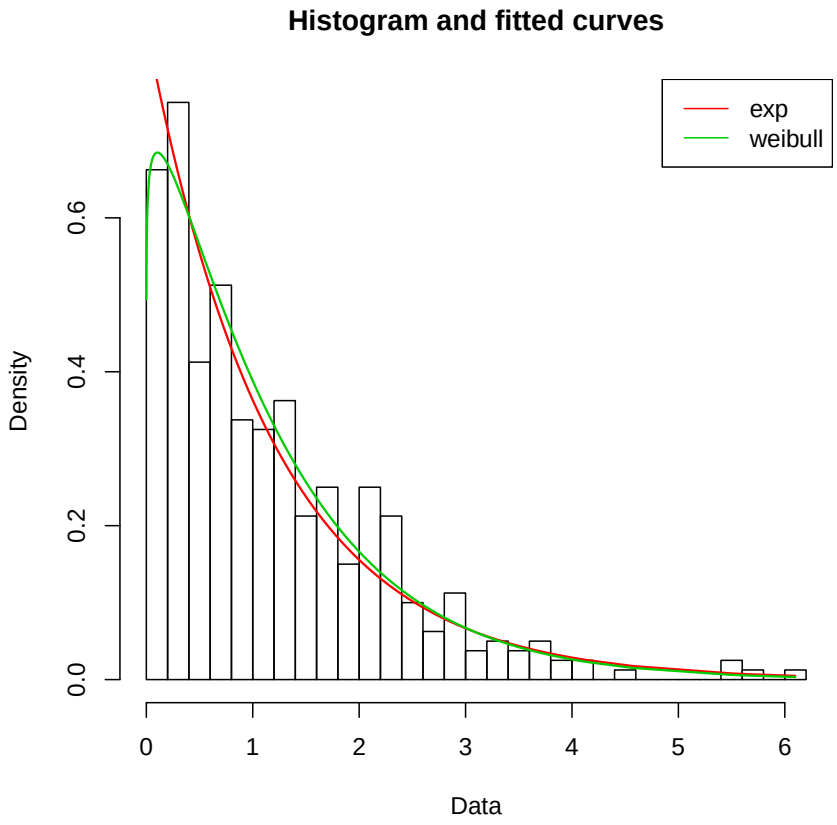
Scoring is clearly the greatest fun of a football match. People extensively interested in football certainly enjoys a well-played match even without many goals. Nevertheless, for the more average viewer of football, scorings are the most important feature of the game. Before a match start one would for sure like to know how many goals there will be scored in order to decide whether to watch it or not. What is the probability that a match ends up rather regular with say 2 goals in total? Or what is the probability of the "booring" match result 0-0, i.e. no goals scored? Considering a person without much knowledge of the teams of a certain match, e.g. for a person randomly switching into a broadcasted Premier League match at lunch time on a Saturday. This example tries to answer what his expectations regarding the number of goals should be like.

We encounter a data set with the counts of the number of goals scored in each Premier League match played in a total of 11 consecutive seasons, starting from the 2001–2002 season and including the 2011–2012 season which finished off in late May. With 20 teams in the league making up a total of 38 rounds with 10 matches each round, the data sets consist of the total number of goals in 4120 matches. We also consider the data set as iid. The assumption of identical distributed data may not hold when considering which teams are playing, but since we here assume that this is unknown in the sense that the viewer does not know much of these teams anyway, this should not matter. Moreover, one may argue that the scoring rate has changed over the years, but since Premier League always has been an attractive league with a high level, it seems reasonable to assume that the distribution of the number of goals has not changed over the years either. The assumption of iid data are thus deemed reasonable.

Since the number of goals scored during a match is a count, it has been a common approach to use the well-known Poisson distribution to model such data. Especially, models for predicting outcomes of football matches based on Poisson models, are discussed in Claeskens and Hjort (2008, example 2.8). Furthermore the Norwegian Computing Center has provided statistical predictions before the European Cups and World Cups the last couple of times arranged. Also in this example modified Poisson models are the main ingredient. We call attention on estimating four different focus parameters. We perform model selection using FIC to select model when the focus parameter is the probability of scoring exactly 0, 1 and 2 goals in a match, one at a time. In the end, we also focus on estimation of the probability of what we will call a "guaranteed fun match", a match with 4 goals or more. In addition to the Poisson model, we propose the so-called CoM–Poisson model developed by Conway and Maxwell in addition to the usual nonparametric model. The CoM–Poisson model has a probability mass function on the form:

$$Pr\{X = k\} = f_{\text{CoM-P}}(k; \lambda, \alpha) = \frac{\lambda^k}{(k!)^\alpha} \frac{1}{Z(\lambda, \alpha)},$$

Figure 3.4: Histogram of the simulated data with the density curves of the fitted exponential and Weibull distributions.



for $k = 0, 1, \ldots$ and parameters $\lambda > 0$ and $\alpha \geq 0$. Here $Z(\lambda, \alpha)$ is the normalization constant given by

$$Z(\lambda, \alpha) = \sum_{k=0}^{\infty} \frac{\lambda^k}{(k!)^\alpha}.$$

When $\alpha = 1$ the CoM–Poisson distribution reduces to the regular Poisson distribution. Even if not of direct interest here, one might prove that when $\alpha \to \infty$, the distribution reduces to a certain Bernoulli distribution. Furthermore, when $\alpha = 0$ and $\lambda < 1$, the distribution reduces to a geometric distribution.

Since all of the focus parameters we encounter in this example are of the smooth form discussed previously, they are all applicable to our FIC scheme. Even if there is only two cases this will make a difference, we will also here be using the adjusted FIC version for all 4 situations.

The FIC tables 3.9, 3.10 and 3.11 below give the results of the model selections when interest is on estimation of the point mass probabilities at 0, 1 and 2. As is apparent from the tables, each of the models are best at estimating one of the probabilities each, i.e. the FIC scheme selects different models for each of the three tasks. Figure 3.5 shows how the data spreads over different total number of goals and how the fitted parametric distributions match the data.

	$\widehat{\mu}$	dim	$\widehat{\text{bias}}^*$	$\widehat{\text{sd}}$	$\widehat{\text{RMSE}}$	Rank
Nonpar	0.0823	Inf	0	0.0043	0.0043	2
Poisson	0.0722	1	0.0094	0.0019	0.0096	3
CoM.poisson	0.0789	2	0.0024	0.0034	0.0042	1

Table 3.9: Results of the adjusted FIC scheme of example 3 for the probability of exactly 0 goals.

	$\widehat{\mu}$	dim	$\widehat{\text{bias}}^*$	$\widehat{\text{sd}}$	$\widehat{\text{RMSE}}$	Rank
Nonpar	0.1809	Inf	0	0.0060	0.0060	1
Poisson	0.1897	1	0.0071	0.0030	0.0077	2
CoM.poisson	0.1914	2	0.0092	0.0030	0.0096	3

Table 3.10: Results of the adjusted FIC scheme of example 3 for the probability of exactly 1 goal.

We see that the CoM–Poisson model is deemed the best at estimating the probability of the match result 0-0 corresponding to no goals. The winning model is just slightly better than the nonparametric model, but these two models are much better than the Poisson model. The estimated probability of a 0-0 under the winning CoM–Poisson model is 7.89%.

On the other hand, the nonparametric model is regarded the best model for estimating the probability of either the score 1-0 or 0-1. The Poisson model is second best and the CoM–Poisson model third best. The estimate of the probability of such an outcome is under the winning nonparametric model 18.09%.

Furthermore, the Poisson model wins according to FIC for the situation when the probability of exactly 2 goals in a certain match is of interest. Second best is the CoM–Poisson model, whereas the nonparametric model is devoted to the third place. As seen from table 3.11 the estimates based on all models are very similar. The winning model gives an estimated probability of 24.94%, i.e. almost every fourth match ends with exactly 2 goals scored. Note also that with this focus, the squared bias of both parametric models is estimated as zero, which are caused by the adjustment of the FIC formula.

A final application of these data, model selection is performed for estimation of a match with many goals, defined as 4 goals or more. Table 3.12 gives a FIC table with these results. All models perform almost equally well at this task in terms of RMSE, and they all have very similar estimates. The simplest Poisson model is however the winning model, with the CoM–Poisson model slightly behind and on a third place the nonparametric model. Under the winning Poisson model the probability of a match ending with at least four goals is estimated to 27.03%. It is for sure pleasing to see that such an event, which often corresponds to a good game of football occurs with such a big probability.

Finally we note that for sure this type of situations could have been made more specific, in terms of who is playing. By only considering matches where a certain team plays, one could estimate the different aspects of the matches for this particular team. The data size is then greatly reduced. For the teams that have been in the upper division all these seasons, we got data for 418 matches. One could also consider estimation of the number of goals or

	$\widehat{\mu}$	dim	$\widehat{\text{bias}}^*$	$\widehat{\text{sd}}$	$\widehat{\text{RMSE}}$	Rank
Nonpar	0.2488	Inf	0	0.0067	0.0067	3
Poisson	0.2494	1	0	0.0015	0.0015	1
Com.poisson	0.2435	2	0	0.0028	0.0028	2

Table 3.11: Results of the adjusted FIC scheme of example 3 for the probability of exactly 2 goals.

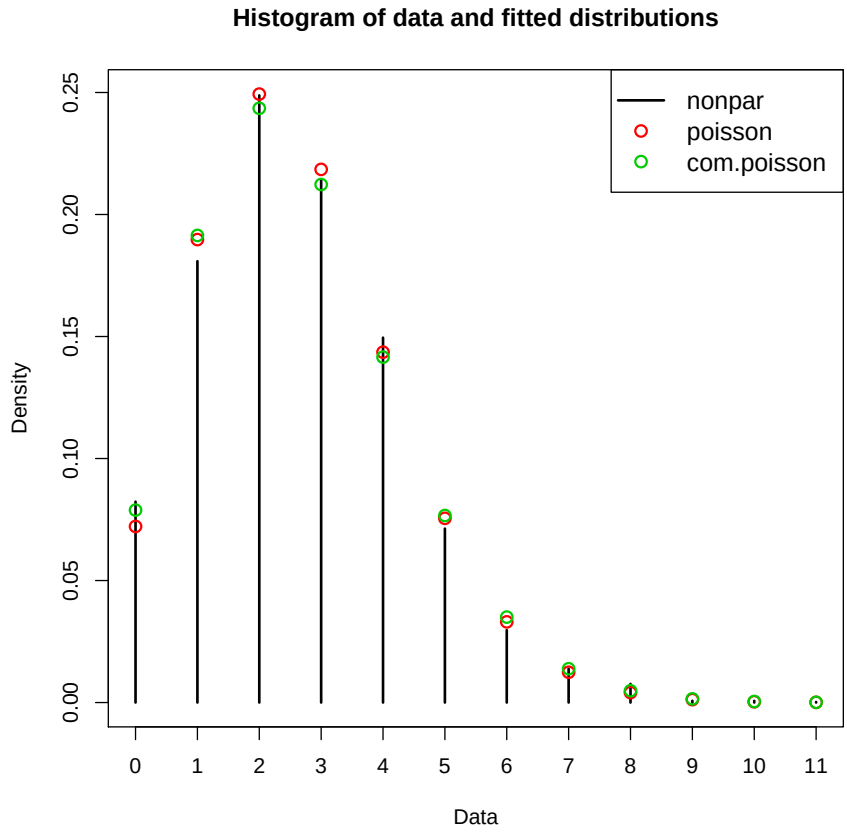
	$\widehat{\mu}$	dim	$\widehat{\text{bias}}^*$	$\widehat{\text{sd}}$	$\widehat{\text{RMSE}}$	Rank
Nonpar	0.2737	Inf	0	0.0069	0.0069	3
Poisson	0.2703	1	0	0.0056	0.0056	1
CoM.poisson	0.2740	2	0	0.0057	0.0057	2

Table 3.12: Results of the adjusted FIC scheme of example 3 for the probability 4 or more goals.

maybe the winning margin in a match between two teams of the so-called "big four", consisting of Manchester United, Chelsea, Arsenal and Liverpool. These teams have had rather steady season performances over these years and the matches between these teams are always exciting. Especially the winning margin may be interesting here, since one might expect that the these matches does not give high winning margins, as opposed to a game against a team in the bottom of the league table.[5] Not to drag this example to much out, we do nevertheless stop out analysis at this point.

[5] A possible outlier may however be Manchester United's 8-2 win over Arsenal in late August 2011.

Figure 3.5: Histogram of the number of goals scored at 4120 Premier League matches, with the probability mass function of the fitted Poisson and CoM–Poisson distributions.



Chapter 4

FIC for censored iid data

Censored data occurs in many situations, and are of special importance when studying survival analysis. The hazard function and the survival function are often of great interest in such studies. To estimate the cumulative hazard and survival function, the nonparametric Nelson–Aalen and Kaplan–Meier estimators are often used. If a certain parametric model fits well to data, these function, one may however also be estimated well by fitting a parametric distribution. To choose which method to rely on, model selection is therefore of interest here as well.

In this chapter this is exactly what we will study, although in a focused sense. We assume that we observe data which originally are iid from some underlying common distribution, but that the data possibly are censored. Since estimation of the cumulative hazard and the survival function at some point are the most interesting foci for these kind of data with well-defined and commonly applied nonparametric estimators, we will mainly focus on these two focus parameters. The aim of the chapter is to derive FIC schemes for these two focus parameters. The basic idea on construction of these schemes is the same as for the regular iid case in chapter 3, i.e. to base the estimators on joint limiting distributions for the μ estimators. First we do however give a short introduction to the basics of stochastic processes, martingales and survival analysis. We then derive the mentioned joint limiting distributions under some assumptions we state, and use the limiting distribution to give FIC scores which form a model selection apparatus. We further discuss the underlying conditions in short terms. Towards the end of the chapter we discuss other focus parameters before we finish off with a simple illustration.

4.1 Stochastic processes and survival analysis

When working with survival analysis where events occur over time and so-called "censoring" of the data is common, it is fruitful to treat the data as a stochastic processes. Martingale theory may then be used to derive properties of these estimators. The basic concepts of the two are introduced below.

4.1.1 Stochastic processes

A stochastic process $X(t)$ is a random variable changing over time (t) often representing the evolution of some random variable over time. A counting process $N(t)$ is a type of stochastic process that counts the number of events having occurred up until time t, for $t \in [0, \tau]$. For the purpose of using counting processes in this thesis, it is clever to attach a so-called intensity

process $\lambda(t)$ to the process. The intensity process is an integrable function loosely defined as the conditional probability of a jump for the counting process in a small interval given the past, divided by the length of the interval. A useful property of counting processes is that when integrating over them, they reduce integrals to sums. I.e. for some function $h(s)$, we have

$$\int h(s)\,\mathrm{d}N(s) = \sum_{s\in A} h(s),$$

where A is the set of jumping times for the counting process $N(t)$.

Counting processes and their intensity processes naturally creates so-called martingales. A stochastic process $M(t)$ is said to be a martingale (or have the martingale property) if the expected value of the process at all future time points given the past, is equal to the current value. It can be shown that the following process actually is a martingale:

$$M(t) = N(t) - \int_0^t \lambda(s)\,\mathrm{d}s.$$

A stochastic process $H(t)$ is said to be predictable if the value of the process at any time t is know just before t. It can in fact be shown that if $H(t)$ is a predictable process, the stochastic integral

$$M^*(t) = \int_0^t H(s)\,\mathrm{d}M(s),$$

also is a martingale.

4.1.2 Survival analysis

The field of survival analysis is of great practical interest. In this field one studies the time until death, occurrence of a disease or failure of a certain component in a mechanical system. These events are studied for a set of "individuals". The term "individual" is used since most often one is working with exactly individuals, and we will also adopt this terminology. In theory one may however consider e.g. components in a system. The data one encounters in such situations are most often incomplete in the way that the event of interest is not observed for all individuals. The reason for this lies in the nature of the data. All individuals in a survey do not need to e.g. get a certain disease, and even if a component would fail at some time point, one cannot wait forever to analyze the data. When the event of interest is not observed for a certain individual, we say that the individual is censored at the time when we stopped the observation. Thus, the data one usually analyzes consist of the time points of what happens first of censoring and the event of interest for the n individuals under observation, along with the information of whether or not each of the individuals were censored. It is exactly this somewhat inconvenient nature of the data that demands a whole field of statistics.

In more mathematical terms, we will work under the following framework: Assume that a set of n individuals of the same population have been under observation for a time period starting at 0 and ending at some time point $\tau < \infty$. Let these individuals have true underlying iid survival times $T_1, \ldots, T_n$, stemming from a continuous distribution function with cdf $G(t)$ and density $g(t)$.[1] For each of these individuals, we only observe $\tilde{T}_i = \min\{T_i, C_i\}$ and D_i, where

[1] We will assume that the data have a continuous distribution in this chapter, since this is clearly the most common situation. Discrete survival times may be of interest in some situation, but since most theory available are connected to continuous data, we will also stick to that assumption.

C_i is the censoring time of individual i and $D_i = \mathbf{1}_{\{\tilde{T}_i = T_i\}}\left(\tilde{T}_i, T_i\right)$ is the indicator of censoring. We also assume that the C_i's are random. Assume also that so-called independent censoring apply, which a bit sloppy means that the censoring rate does not change over time. The exact definition of independent censoring is given e.g. in Aalen et al. (2008, page 30). Furthermore, we let $N_i(t)$ be the counting process indicating whether the single event of interest has occurred or not for individual i. Thus, $N_i(t)$ is a counting process taking only 0 or 1. In addition we assume that the multiplicative intensity model holds, i.e. that this counting process has an intensity process $\lambda_i(t) = \alpha(t)Y_i(t)$, where $Y_i(t) = \mathbf{1}_{\{\tilde{T}_i \geq t\}}(t)$ is the indicator of individual i still being "at risk" at time t, and $\alpha(t)$ is the unknown hazard function common for the population.

For an absolute continuous distribution, which we mostly will be working with, the hazard function is defined as $\alpha(t) = g(t)/S(t)$, for $g(t)$ the usual true density and $S(t) = 1 - G(t)$. When deriving results later on it will be convenient to accumulate the counting process into

$$N(t) = \sum_{i=1}^{n} N_i(t),$$

which has intensity process

$$\lambda(t) = \sum_{i=1}^{n} \lambda_i(t) = \alpha(t)Y(t),$$

where $Y(t) = \sum_{i=1}^{n} Y_i(t)$ is the total number of individuals at risk at time t. For later convenience, we also define $L(t) = \mathbf{1}_{\{Y(t)>0\}}(t)$.[2] Finally we define

$$M(t) = N(t) - \int \lambda(s)\,\mathrm{d}s.$$

It is common practice to omit the subscript n for these processes even if they depend on the sample size n, and we will therefore also follow this practice. With these quantities we define the nonparametric estimator of the cumulative hazard function $A(t) = \int_0^t \alpha(s)\,\mathrm{d}s$ as

$$\widehat{A}_{\mathrm{np}}(t) = \sum_{T_i \leq t} \frac{1}{Y(T_i)} = \int_0^t \frac{L(s)}{Y(s)}\,\mathrm{d}N(s),$$

and the nonparametric estimator for the survival function $S(t) = 1 - G(t) = Pr\{T_i > t\}$ as

$$\widehat{S}_{\mathrm{np}}(t) = \prod_{T_i \leq t} \left(1 - \frac{1}{Y(T_i)}\right).$$

Another way to model the data, is to fit a parametric model. As for the regular iid situation, we will focus on maximum likelihood estimation. Since we now deal with censored data, the likelihood of the data and will be different. Using the processes defined above, we may write the likelihood as

$$L_n(\theta) = \exp\left(\int_0^\tau \log(\alpha(s;\theta))\,\mathrm{d}N(s) - Y(s)\alpha(s;\theta)\,\mathrm{d}s\right), \tag{4.1}$$

for θ the parameter vector of the parametric distribution with hazard function $\alpha(t;\theta)$, density $f(t;\theta)$ and cdf $F(t;\theta)$, in the situation of random censoring. The ML estimator $\widehat{\theta}_n$ is as usual

[2] The convention is to use $J(s)$ for this, but to avoid confusion, we use $L(s)$ to denote this quantity instead.

defined as the maximizer of equation (4.1), and aiming at estimating some least false parameter vector θ_0. In this situation θ_0 will be defined as the minimizer of the generalized Kullback–Leibler divergence

$$\int_0^\tau y(s)\left(\alpha(s)\left(\frac{\log(\alpha(s))}{\log(\alpha(s;\theta))}\right)-(\alpha(s)-\alpha(s;\theta))\right)\mathrm{d}s,$$

where $y(s)$ is a nonnegative function having the property that $|Y(s)/n - y(s)| \xrightarrow{P} 0$ uniformly for all s. Equivalently θ_0 is defined as the maximizer of

$$\int_0^\tau y(s)(\alpha(s)\log(\alpha(s;\theta))-\alpha(s;\theta))\,\mathrm{d}s,$$

which one may show (see e.g. Hjort (1992)) is the limit of $(1/n)L_n(\theta)$ with probability 1. As a result, the parametric estimators of respectively the cumulative hazard rate and the survival function are created by simply inserting $\widehat{\theta}_n$ for θ_0 in the least false parametric analogues $A(t;\theta_0)$ and $S(t;\theta_0)$.

As mentioned, the hazard function and survival probability is often of main interest in such studies. It is very common to use these estimators to investigate these properties of the distribution, mainly because it is often hard to know anything about the underlying distribution in advance of such studies. However, if a certain parametric distribution really is true or is fairly close to the true distribution of the survival times, parametric models will do a better job for estimation purposes. This should motivate the FIC derivations of the remaining chapter.

4.2 Limiting distribution

As mentioned in the introduction of this chapter, counting processes and martingales play a central role in modern survival analysis. Most derivations and properties are dealt with by treating the data and problems in terms of suitable defined processes of this type. A wide range of limit theory and useful formulations are available within this branch, making it an optimal framework for theoretical treatment of censored data in this context.

We will now state and prove a lemma including the joint limiting distribution of the non-parametric and parametric estimators for the cumulative hazard at some time point t. Since the marginal limiting distributions of both these estimators have been carried out before, there is no need to start entirely from the bottom once again. We will base our derivations on (Hjort, 1992, Theorem 2.1), and its proof which gives the limiting distribution of the ML estimator of the parametric distribution without assuming the parametric model is correct. This derivation will be mixed with Andersen et al. (1993, Theorem IV.1.2) and pieces of its proof. Before we state and prove the lemma we are going to define a few helpful quantities and state regularity conditions which we will assume for the situation we are working within. To ease the comparison with the regular iid situation we will use the same notation as in chapter 3, except that

we include a $'$ at the end of the quantities to distinguish the two types.

$$\begin{aligned}
J' &= \int_0^\tau y(s)\left[\psi(s;\theta_0)\psi(s;\theta_0)^t\alpha(s;\theta_0) - \dot{\psi}(s;\theta_0)(\alpha(s)-\alpha(s;\theta_0))\right]\,\mathrm{d}s,\\
K' &= \int_0^\tau \left[y(s)\psi(s;\theta_0)\psi(s;\theta_0)^t\alpha(s) + \left(\psi(s;\theta_0)\iota(s)^t + \iota(s)\psi(s;\theta_0)^t\right)\alpha(s;\theta_0)\right]\,\mathrm{d}s,\\
\nu'(t) &= \int_0^t \frac{\alpha(s)}{y(s)}\,\mathrm{d}s,\\
Q'(t) &= \int_0^t \psi(s;\theta_0)\alpha(s)\,\mathrm{d}s - \int_0^\tau y(s)\psi(s;\theta_0)(\alpha(s)-\alpha(s;\theta_0))\nu'(\max\{t,s\})\,\mathrm{d}s,
\end{aligned}$$

where

$$\begin{aligned}
\psi(s;\theta) &= \frac{\partial}{\partial\theta}\log(\alpha(s;\theta)),\\
\dot{\psi}(s;\theta) &= \frac{\partial^2}{\partial\theta\partial\theta^t}\log(\alpha(s;\theta)),\\
\iota(s) &= \int_0^s y(u)\psi(u;\theta_0)(\alpha(u)-\alpha(u;\theta_0))\,\mathrm{d}u,
\end{aligned}$$

and $y(s)$ some nonnegative function. In addition, the key quantity $\overline{U}_n = \frac{\partial}{\partial\theta}(1/n)\log(L_n(\theta_0))$ may be written as

$$\begin{aligned}
\overline{U}_n &= \frac{1}{n}\int_0^\tau \psi(s;\theta_0)(\,\mathrm{d}N(s) - Y(s)\alpha(s;\theta_0))\,\mathrm{d}s\\
&= \frac{1}{n}\int_0^\tau \psi(s;\theta_0)(\,\mathrm{d}M(s) + Y(s)(\alpha(s)-\alpha(s;\theta_0))\,\mathrm{d}s).
\end{aligned}$$

We now state assumptions which we will be working under in what follows.

Assumption 4.2.1. *For some $t \in (0,\tau)$, we assume that $y(s)$ has the property that*

$$\inf_{s\in[0,t]} y(s) > 0, \tag{4.2}$$

and

$$\sup_{s\in[0,t]}\left|\frac{Y(s)}{n} - y(s)\right| \xrightarrow{P} 0. \tag{4.3}$$

$$n\int_0^t \frac{L(s)}{Y(s)}\alpha(s)\mathbf{1}_{\{|\sqrt{n}L(s)/Y(s)|>\epsilon\}}(s)\,\mathrm{d}s \xrightarrow{P} 0, \tag{4.4}$$

and that $\alpha(s)/y(s)$ is integrable on $[0,t]$. In addition, for a p-dimensional parameter vector θ of the parametric familiy of distributions with cdf F_θ and least false parameter θ_0, we assume that θ_0 is unique, that the hazard functions $\alpha(s)$ and $\alpha(s;\theta_0)$ are both bounded away from zero as s runs from 0 to τ, and that

$$\widehat{\theta}_n = \theta_0 + J'^{-1}\overline{U}_n + o_p\left(\frac{1}{\sqrt{n}}\right). \tag{4.5}$$

Assume also that

$$\left.\frac{\partial F(t;\theta)}{\partial\theta}\right|_{\theta_0} \neq 0.$$

We are now ready to give the key lemma. Note that even if the limiting variance and covariance terms depend on t, we will for ease of representation omit the additional (t).

Lemma 4.2.2. *When the relations and conditions of assumptions 4.2.1 hold, the following limiting distribution appears:*

$$\sqrt{n}\begin{pmatrix}\widehat{A}_{\text{np}}(t) - A_{\text{true}}(t)\\ \widehat{A}_{\text{pm}}(t) - A_{0,\text{pm}}(t)\end{pmatrix} \xrightarrow{L} N_2\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}V'_{A,\text{np}} & V'_{A,\text{pm,np}}\\ V'_{A,\text{pm,np}} & V'_{A,\text{pm}}\end{pmatrix}\right),$$

where

$$\begin{aligned}
V'_{A,\text{np}} &= \nu'(t),\\
V'_{A,\text{pm}} &= \left(\frac{\partial A(t;\theta)}{\partial\theta}\Big|_{\theta=\theta_0}\right)^t J'^{-1}K'J'^{-1}\left(\frac{\partial A(t;\theta)}{\partial\theta}\Big|_{\theta=\theta_0}\right),\\
V'_{A,\text{pm,np}} &= \left(\frac{\partial A(t;\theta)}{\partial\theta}\Big|_{\theta=\theta_0}\right)^t J'^{-1}Q'(t).
\end{aligned}$$

Note that we have denoted the true cumulative hazard at the point t by $A_{\text{true}}(t)$ in the above lemma. This is done to make it as similar to the corresponding lemma 3.1.2 for the regular iid situation as possible. There is no difference between this quantity and what we earlier have denoted simply $A(t)$.

Proof. First, assume that we have shown the following limiting distribution:

$$\sqrt{n}\begin{pmatrix}\widehat{A}_{\text{np}}(t) - A_{\text{true}}(t)\\ \widehat{\theta}_n - \theta_0\end{pmatrix} \xrightarrow{L} N_{p+1}\left(0, \Sigma'\right), \tag{4.6}$$

where Σ' may be written as a block matrix of the form

$$\Sigma' = \begin{pmatrix}\Sigma'_{00} & \Sigma'_{01}\\ \Sigma'_{10} & \Sigma'_{11}\end{pmatrix},$$

and

$$\begin{aligned}
\Sigma'_{00} &= J'^{-1}K'J'^{-1},\\
\Sigma'_{11} &= \nu',\\
\Sigma'_{01} &= \left(\Sigma'_{10}\right)^t = J'^{-1}Q'(t).
\end{aligned}$$

We then apply the delta method (theorem B.2.8) with the following transformation function:

$$S_A(z,x) = \begin{pmatrix}z\\ A(t;x)\end{pmatrix}.$$

The function has Jacobian matrix given by

$$\dot{S}_A(z,x) = \begin{pmatrix}1 & 0\\ 0 & \frac{\partial A(t;x)}{\partial x}\end{pmatrix}.$$

Thus, the delta method gives

$$\begin{aligned}\sqrt{n}\begin{pmatrix}\widehat{A}_{\text{np}}(t)-A_{\text{true}}(t)\\ A(t;\widehat{\theta}_n)-A(t;\theta_0)\end{pmatrix} &= \sqrt{n}\begin{pmatrix}\widehat{A}_{\text{np}}(t)-A_{\text{true}}(t)\\ \widehat{A}_{\text{pm}}(t)-A_{0,\text{pm}}(t)\end{pmatrix}\\ &\xrightarrow{L} N_2\left(0,\left(\dot{S}_A(A_{\text{true}},\theta_0)\right)^t\Sigma'\left(\dot{S}_A(A_{\text{true}},\theta_0)\right)\right)\\ &= N_2\left(\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}V'_{A,\text{np}} & V'_{A,\text{pm,np}}\\ V'_{A,\text{pm,np}} & V'_{A,\text{pm}}\end{pmatrix}\right),\end{aligned}$$

which is exactly the limit result we should prove. What remains now is to validate the first limiting distribution. To do that we first introduce the quantity

$$A^*(t)=\int_0^t L(s)\alpha(s)\,\mathrm{d}s.$$

From assumption 4.2.1, and specifically equations (4.2) and (4.3) we get by Andersen et al. (1993, theorem IV.1.2 and succeeding comment) that

$$\sqrt{n}(A^*(t)-A(t))\xrightarrow{P}0.$$

We also note that

$$\widehat{A}_{\text{np}}(t)-A^*(t)=\int_0^t\frac{L(s)}{Y(s)}[\,\mathrm{d}N(s)-Y(s)\alpha(s)\,\mathrm{d}s]=\int_0^t\frac{L(s)}{Y(s)}\,\mathrm{d}M(s)=\int_0^\tau\frac{L'(s)}{Y(s)}\,\mathrm{d}M(s),$$

where $L'(s)=L(s)\mathbf{1}_{\{s\le t\}}(s)$. Using the two latest results in addition to the assumed key relation (4.5) for the parametric distribution, we get

$$\begin{aligned}\sqrt{n}\begin{pmatrix}\widehat{A}_{\text{np}}(t)-A_{\text{true}}(t)\\ \widehat{A}_{\text{pm}}(t)-A_{0,\text{pm}}(t)\end{pmatrix} &= \sqrt{n}\begin{pmatrix}\int_0^\tau\frac{L'(s)}{Y(s)}\,\mathrm{d}M(s)+o_p\left(\frac{1}{\sqrt{n}}\right)\\ J'^{-1}\overline{U}_n+o_p\left(\frac{1}{\sqrt{n}}\right)\end{pmatrix}\\ &= \begin{pmatrix}1&0\\0&J'^{-1}\end{pmatrix}\sqrt{n}\frac{1}{n}\int_0^\tau\begin{pmatrix}\frac{L'(s)}{Y(s)/n}\\ \psi(s;\theta_0)\end{pmatrix}\Bigg[\mathrm{d}M(s)\\ &\quad+\begin{pmatrix}0\\e_p\end{pmatrix}Y(s)(\alpha(s)-\alpha(s;\theta_0))\,\mathrm{d}s\Bigg]+\begin{pmatrix}o_p(1)\\o_p(1)\end{pmatrix}\\ &= \begin{pmatrix}1&0\\0&J'^{-1}\end{pmatrix}\int_0^\tau\begin{pmatrix}\frac{L'(s)}{Y(s)/n}\\ \psi(s;\theta_0)\end{pmatrix}\Bigg[\mathrm{d}\frac{M(s)}{\sqrt{n}}\\ &\quad+\begin{pmatrix}0\\e_p\end{pmatrix}\sqrt{n}\left(\frac{Y(s)}{n}-y(s)\right)(\alpha(s)-\alpha(s;\theta_0))\,\mathrm{d}s\Bigg]+\begin{pmatrix}o_p(1)\\o_p(1)\end{pmatrix}\\ &= \begin{pmatrix}1&0\\0&J'^{-1}\end{pmatrix}B_n+o_p(1),\end{aligned}\tag{4.7}$$

where e_p is a p-dimensional vector with only 1's. In the main expression three lines above we have used the fact that $\int_0^\tau\psi(s;\theta_0)y(s)(\alpha(s)-\alpha(s;\theta_0))\,\mathrm{d}s=0$ since this expression is just the derivative of the function minimizing the Kullback–Leibler divergence. Therefore, this expression could be added without any adjustments. We now further investigate B_n of equation

(4.7) above by using martingale theory and other useful results presented in Hjort (1992). M is a martingale based on a certain counting process, and by the general theory of Andersen et al. (1993), including a slightly stronger version of theorem B.2.12, we get that $D_n = M/\sqrt{n}$ has limiting process which coincides with a zero-mean Gaussian process[3] D, where

$$\mathrm{Var}(dD_n(s)) = \mathrm{Var}(dD(s)) = y(s)\alpha(s)\,\mathrm{d}s.$$

Furthermore, $Z_n = \sqrt{n}(Y/n - y)$ also converges to a Gaussian zero-mean process Z, this one with

$$\mathrm{Cov}(Z_n(s), Z_n(t)) = \mathrm{Cov}(Z(s), Z(t)) = y(\max\{s,t\}) - y(s)y(t),$$

see Hjort (1992). Furthermore, in the same paper the author shows that

$$\mathrm{Cov}(\,\mathrm{d}D_n(s), Z_n(t)) = \mathrm{Cov}(\,\mathrm{d}D(s), Z(t)) = -\alpha(s)\,\mathrm{d}s\; y(t)\mathbf{1}_{\{s<t\}}(s,t).$$

Finally, it is stated that $(D_n, Z_n) \xrightarrow{L} (D, Z)$. Observe that from condition (4.4), we get

$$\begin{aligned} n\int_0^s \frac{L(u)}{Y(u)}\alpha(u)\,\mathrm{d}u &= \int_0^s \frac{\mathbf{1}_{\{Y(u)>0\}}(u)}{Y(u)/n}\alpha(u)\,\mathrm{d}u \\ &\xrightarrow{P} \int_0^s \frac{\mathbf{1}_{\{y(u)>0\}}(u)}{y(u)}\alpha(u)\,\mathrm{d}u \\ &= \int_0^s \frac{\alpha(u)}{y(u)}\,\mathrm{d}u = \nu'(s). \end{aligned}$$

The first part of Andersen et al. (1993, theorem IV.1.2) assures that the nonparametric part of B_n may be applied to the martingale central limit (theorem B.2.12). Combining this with function space asymptotics from Billingsley (1999) and Andersen and Borgan (1985), which assures that the parametric part of B_n may also be handled by the martingale CLT (see Hjort (1992)), we get that

$$B_n \xrightarrow{L} B \overset{d.}{=} \int_0^\tau \begin{pmatrix} \mathbf{1}_{\{s\le t\}}(s)/y(s) \\ \psi(s;\theta_0) \end{pmatrix} [dD(s) + \begin{pmatrix} 0 \\ e_p \end{pmatrix} Z(s)(\alpha(s) - \alpha(s;\theta_0))\,\mathrm{d}s] = \begin{pmatrix} \mathrm{one}^* \\ \mathrm{one} \end{pmatrix} + \begin{pmatrix} 0 \\ \mathrm{two} \end{pmatrix},$$

by using the one/two notation of Hjort (1992). Expressions for the variance and covariance of one and two are already given in the paper, and it is shown that $\mathrm{Var}(one + two) = K'$. Thus, all we need to do is to verify the expressions for $\mathrm{Var}(\mathrm{one}^*)$, $\mathrm{Cov}(\mathrm{one}^*, \mathrm{one})$ and $\mathrm{Cov}(\mathrm{one}^*, \mathrm{two})$. Using properties of stochastic integrals as in Hjort (1992), we get

$$\mathrm{Var}(\mathrm{one}^*) = \int_0^\tau \frac{\mathbf{1}_{\{s\le t\}}(s)}{y(s)^2} y(s)\alpha(s)\,\mathrm{d}s = \int_0^t \frac{\alpha(s)}{y(s)}\,\mathrm{d}s = \nu'(t).$$

Moreover, we get

$$\mathrm{Cov}(\mathrm{one}^*, \mathrm{one}) = E[\mathrm{one}^*\mathrm{one}] = \int_0^\tau \frac{\mathbf{1}_{\{s\le t\}}(s)}{y(s)} \psi(s;\theta_0) y(s)\alpha(s)\,\mathrm{d}s = \int_0^t \psi(s;\theta_0)\alpha(s)\,\mathrm{d}s,$$

[3] A Gaussian process is a stochastic process whose realizations are normally distributed random variables.

and

$$\begin{aligned}
\mathrm{Cov}(\mathrm{one}^*, \mathrm{two}) &= E[\mathrm{one}^*\mathrm{two}] \\
&= \int_0^\tau \int_0^\tau \frac{\mathbf{1}_{\{u\le t\}}(u)}{y(u)} \psi(v;\theta_0)(\alpha(v)-\alpha(v;\theta_0))(-\alpha(u)y(v)\mathbf{1}_{\{u\le v\}}(u,v))\,\mathrm{d}u\,\mathrm{d}v \\
&= -\int_0^\tau \int_0^{\min\{t,v\}} y(v)\psi(v;\theta_0)(\alpha(v)-\alpha(v;\theta_0))\frac{\alpha(u)}{y(u)}\,\mathrm{d}u\,\mathrm{d}v \\
&= -\int_0^\tau y(v)\psi(v;\theta_0)(\alpha(v)-\alpha(v;\theta_0))\int_0^{\min\{t,v\}} \frac{\alpha(u)}{y(u)}\,\mathrm{d}u\,\mathrm{d}v \\
&= -\int_0^\tau y(v)\psi(v;\theta_0)(\alpha(v)-\alpha(v;\theta_0))\nu'(\min\{t,v\})\,\mathrm{d}v.
\end{aligned}$$

As a result, we end up with

$$\mathrm{Cov}(\mathrm{one}^*, \mathrm{one}+\mathrm{two}) = Q'(t).$$

Consequently we get that covariance matrix of O is given by

$$\begin{pmatrix} \nu'(t) & Q'(t) \\ Q'(t) & K' \end{pmatrix}.$$

Thus matrix multiplication with $\begin{pmatrix} 1 & 0 \\ 0 & J'^{-1} \end{pmatrix}$ and the use of Slutsky's theorem (B.2.6) gives exactly relation (4.6), and the proof is completed. □

In the above lemma $\frac{\partial A(t;\theta)}{\partial\theta}$ may be simplified to $\frac{\frac{\partial}{\partial\theta}F(t;\theta)}{1-F(t;\theta)}$ whenever f is absolutely continuous, since we then have that

$$\begin{aligned}
A(t;\theta) &= \int_0^t \frac{f(s;\theta)}{1-F(s;\theta)}\,\mathrm{d}s = -\int_0^t \frac{\partial}{\partial s}\log(1-F(s;\theta))\,\mathrm{d}s \\
&= -\log(1-F(t;\theta)) + \log(1-F(0;\theta)) \\
&= -\log(1-F(t;\theta)),
\end{aligned}$$

whenever $F(0;\theta) = 0$, which should be the case for parametric models fitted to these data because the survival times are positive.

Lemma 4.2.3. *When the relations and conditions of assumption 4.2.1 hold, the following limiting distribution appears:*

$$\sqrt{n}\begin{pmatrix} \widehat{S}_{\mathrm{np}}(t) - S_{\mathrm{true}}(t) \\ \widehat{S}_{\mathrm{pm}}(t) - S_{0,\mathrm{pm}}(t) \end{pmatrix} \xrightarrow{L} N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} V'_{S,\mathrm{np}} & V'_{S,\mathrm{pm,np}} \\ V'_{S,\mathrm{pm,np}} & V'_{S,\mathrm{pm}} \end{pmatrix}\right),$$

where

$$\begin{aligned}
V'_{S,\mathrm{pm}} &= \left(\frac{\partial S(t;\theta)}{\partial\theta}\Big|_{\theta=\theta_0}\right)^t J'^{-1}K'J'^{-1}\left(\frac{\partial S(t;\theta)}{\partial\theta}\Big|_{\theta=\theta_0}\right), \\
V'_{S,\mathrm{np}} &= \nu'(t)S_{\mathrm{true}}(t)^2, \\
V'_{S,\mathrm{pm,np}} &= -S_{\mathrm{true}}(t)\left(\frac{\partial S(t;\theta)}{\partial\theta}\Big|_{\theta=\theta_0}\right)^t J'^{-1}Q'(t).
\end{aligned}$$

Proof. Since the assumptions are the same as for lemma 4.2.2, those results and the intermediate results of the proof of the lemma, holds also in this situation. Thus, expression (4.6) gives the limiting distribution of

$$\sqrt{n}\begin{pmatrix}\widehat{A}_{\text{np}}(t) - A_{\text{true}}(t)\\ \widehat{\theta}_n - \theta_0\end{pmatrix}. \tag{4.8}$$

By using arguments of the functional delta method joining the cumulative hazard and the survival function through the product integral representation of the survival function (see Andersen et al. (1993, Theorem IV.3.2)), we get asymptotic equivalence between $\sqrt{n}(\widehat{S}_{\text{np}}(t) - S_{\text{true}}(t))$ and $-S_{\text{true}}(t)\sqrt{n}(\widehat{A}_{\text{np}}(t) - A_{\text{true}}(t))$. Thus

$$\begin{aligned}\sqrt{n}\begin{pmatrix}\widehat{S}_{\text{np}}(t) - S_{\text{true}}(t)\\ \widehat{\theta}_n - \theta_0\end{pmatrix} &\overset{eq.}{\sim} \begin{pmatrix}-S_{\text{true}}(t) & 0\\ 0 & 1\end{pmatrix}\sqrt{n}\begin{pmatrix}\widehat{A}_{\text{np}}(t) - A_{\text{true}}(t)\\ \widehat{\theta}_n - \theta_0\end{pmatrix}\\ &\xrightarrow{L} \begin{pmatrix}-S_{\text{true}}(t) & 0\\ 0 & 1\end{pmatrix} B,\end{aligned} \tag{4.9}$$

where $\overset{eq.}{\sim}$ denotes asymptotic equivalence between two distributions, and B is defined as in lemma 4.2.2. Therefore, the limiting distribution also holds for the first expression. Now, the delta method (theorem B.2.8) may be applied to this relation with the following transformation function

$$S_S(z, x) = \begin{pmatrix} z\\ S(t; x)\end{pmatrix}.$$

The function has the Jacobian matrix

$$\dot{S}_S(z, x) = \begin{pmatrix}1 & 0\\ 0 & \frac{\partial S(t;x)}{\partial x}\end{pmatrix}.$$

Consequently, the delta method gives

$$\begin{aligned}\sqrt{n}\begin{pmatrix}\widehat{S}_{\text{np}}(t) - S_{\text{true}}(t)\\ S(t;\widehat{\theta}_n) - S(t;\theta_0)\end{pmatrix} &= \sqrt{n}\begin{pmatrix}\widehat{S}_{\text{np}}(t) - S_{\text{true}}(t)\\ \widehat{S}_{\text{pm}}(t) - S_{0,\text{pm}}(t)\end{pmatrix}\\ &\xrightarrow{L} \begin{pmatrix}-S_{\text{true}}(t) & 0\\ 0 & 1\end{pmatrix}\begin{pmatrix}1 & 0\\ 0 & \frac{\partial S(t;x)}{\partial x}\end{pmatrix} B = \begin{pmatrix}-S_{\text{true}}(t) & 0\\ 0 & \frac{\partial S(t;x)}{\partial x}\end{pmatrix} B\\ &\overset{d.}{=} N_2\left(\begin{pmatrix}0\\ 0\end{pmatrix}, \begin{pmatrix}V'_{S,\text{np}} & V'_{S,\text{pm,np}}\\ V'_{S,\text{pm,np}} & V'_{S,\text{pm}}\end{pmatrix}\right),\end{aligned}$$

which is the limit result we should prove. □

As in the previous chapter, we now provide corollaries with the marginal limiting distributions we will apply directly in later sections.

Corollary 4.2.4. *When the assumption 4.2.1 holds we get the following limiting distributions*

$$\begin{aligned}\sqrt{n}(\widehat{A}_{\text{np}}(t) - A_{\text{true}}(t)) &\xrightarrow{L} N(0, V'_{A,\text{np}}),\\ \sqrt{n}(\widehat{A}_{\text{pm}}(t) - A_{0,\text{pm}}(t)) &\xrightarrow{L} N(0, V'_{A,\text{pm}}),\\ \sqrt{n}(\widehat{b}_A(t) - b_A(t)) &\xrightarrow{L} N(0, V'_{A,b}),\end{aligned}$$

where

$$\begin{aligned}\widehat{b}_A(t) &= \widehat{A}_{\text{pm}}(t) - \widehat{A}_{\text{np}}(t),\\ b_A(t) &= A_{0,\text{pm}}(t) - A_{\text{true}}(t),\\ V'_{A,b} &= V'_{A,\text{pm}} + V'_{A,\text{np}} - 2V'_{A,\text{pmnp}}.\end{aligned}$$

Proof. The proof is exactly the same as for corollary 3.1.3, except that the limiting distribution it is based on is that of lemma 4.2.2 instead. □

The following corollary gives the corresponding marginal limiting distributions for the survival function:

Corollary 4.2.5. *When the assumption 4.2.1 holds we get the following limiting distributions*

$$\begin{aligned}\sqrt{n}(\widehat{S}_{\text{np}}(t) - S_{\text{true}}(t)) &\xrightarrow{L} N(0, V'_{S,\text{np}}),\\ \sqrt{n}(\widehat{S}_{\text{pm}}(t) - S_{0,\text{pm}}(t)) &\xrightarrow{L} N(0, V'_{S,\text{pm}}),\\ \sqrt{n}(\widehat{b}_S(t) - b_S(t)) &\xrightarrow{L} N(0, V'_{S,b}),\end{aligned} \tag{4.10}$$

where

$$\begin{aligned}\widehat{b}_S(t) &= \widehat{S}_{\text{pm}}(t) - \widehat{S}_{\text{np}}(t),\\ b_S(t) &= S_{0,\text{pm}}(t) - S_{\text{true}}(t),\\ V'_{S,b} &= V'_{S,\text{pm}} + V'_{S,\text{np}} - 2V'_{S,\text{pmnp}}.\end{aligned}$$

Proof. The proof is exactly the same as for corollary 3.1.3, except that limiting distribution it is based on is that of lemma 4.2.2 instead. □

4.3 Mean squared error estimators

Before we head over to the actual expressions for the mse estimators, let us state general estimators for each of the quantities in the limiting distributions derived in the previous section. At all points where θ_0 is represented in an otherwise known function, we insert $\widehat{\theta}_n$. Furthermore, since $\widehat{A}_{\text{np}}(t) = \int_0^t \mathrm{d}N(s)/Y(s)$ is an asymptotically unbiased estimator of $\int_0^t \alpha(s)\,\mathrm{d}s$, $\alpha(s)\,\mathrm{d}s$ will be estimated by $\mathrm{d}N(s)/Y(s)$. In addition $y(s)$ will be estimated by $Y(s)/n$. Doing this leads to the following approximations and estimators based on data:

$$
\begin{aligned}
\widehat{V'}_{A,\mathrm{pm}} &= \left(\frac{\partial A(t;\theta)}{\partial\theta}\Big|_{\theta=\widehat{\theta}_n}\right)^t \widehat{J'}^{-1}\widehat{K'}\widehat{J'}^{-1}\left(\frac{\partial A(t;\theta)}{\partial\theta}\Big|_{\theta=\widehat{\theta}_n}\right),\\
\widehat{V'}_{A,\mathrm{np}} &= \widehat{\nu'}(t),\\
\widehat{V'}_{A,\mathrm{pm,np}} &= \left(\frac{\partial A(t;\theta)}{\partial\theta}\Big|_{\theta=\widehat{\theta}_n}\right)^t \widehat{J'}^{-1}\widehat{Q'}(t),\\
\widehat{V'}_{A,b}(t) &= \widehat{V'}_{A,\mathrm{pm}}(t) + \widehat{V'}_{A,\mathrm{np}}(t) - 2\widehat{V'}_{A,\mathrm{pm,np}}(t),\\
\widehat{V'}_{S,\mathrm{pm}} &= \left(\frac{\partial S(t;\theta)}{\partial\theta}\Big|_{\theta=\widehat{\theta}_n}\right)^t \widehat{J'}^{-1}\widehat{K'}\widehat{J'}^{-1}\left(\frac{\partial S(t;\theta)}{\partial\theta}\Big|_{\theta=\widehat{\theta}_n}\right),\\
\widehat{V'}_{S,\mathrm{np}} &= \widehat{\nu'}(t)\widehat{S}_{\mathrm{np}}(t)^2,\\
\widehat{V'}_{S,\mathrm{pmnp}} &= -\widehat{S}_{\mathrm{np}}(t)\left(\frac{\partial S(t;\theta)}{\partial\theta}\Big|_{\theta=\widehat{\theta}_n}\right)^t \widehat{J'}^{-1}\widehat{Q'}(t).
\end{aligned}
$$

where

$$
\begin{aligned}
\widehat{J'} &= \int_0^\tau \frac{Y(s)}{n}\psi(s;\widehat{\theta}_n)\psi(s;\widehat{\theta}_n)^t\alpha(s;\widehat{\theta}_n)\,\mathrm{d}s + \int_0^\tau \frac{Y(s)}{n}\dot{\psi}(s;\widehat{\theta}_n)\alpha(s;\widehat{\theta}_n)\,\mathrm{d}s\\
&\quad - \frac{1}{n}\sum_{T_i\le\tau}\dot{\psi}(T_i;\widehat{\theta}_n),\\
\widehat{K'} &= \frac{1}{n}\sum_{T_i\le\tau}\psi(T_i;\widehat{\theta}_n)\psi(T_i;\widehat{\theta}_n)^t\\
&\quad + \int_0^\tau\Big(\psi(s;\widehat{\theta}_n)\widehat{\iota}(s)^t + \widehat{\iota}(s)\psi(s;\widehat{\theta}_n)^t\Big)\alpha(s;\widehat{\theta}_n)\,\mathrm{d}s,\\
\widehat{\nu'}(t) &= \sum_{T_i\le t}\frac{n}{Y(s)^2},\\
\widehat{Q'}(t) &= \sum_{T_i\le t}\frac{\psi(T_i;\widehat{\theta}_n)}{Y(T_i)} - \frac{1}{n}\sum_{T_i\le\tau}\psi(T_i;\widehat{\theta}_n)\widehat{\nu'}(\min\{t,T_i\})\\
&\quad + \frac{1}{n}\int_0^\tau Y(s)\psi(s;\widehat{\theta}_n)\alpha(s;\widehat{\theta}_n)\widehat{\nu'}(\min\{t,s\})\,\mathrm{d}s,\\
\widehat{\iota'}(s) &= \frac{1}{n}\sum_{T_i\le s}\psi(T_i;\widehat{\theta}_n) - \frac{1}{n}\int_0^s Y(u)\psi(u;\widehat{\theta}_n)\alpha(u;\widehat{\theta}_n)\,\mathrm{d}u.
\end{aligned}
$$

Alternatively $\widehat{K'}$ can be written in a numerically more convenient way (Hjort (1992)) as

$$
\widehat{K'} = \frac{1}{n}\sum_{i=1}^n(\psi(T_i;\widehat{\theta}_n)D_i - A^d(T_i;\widehat{\theta}_n))(\psi(T_i;\widehat{\theta}_n)D_i - A^d(T_i;\widehat{\theta}_n))^t,
$$

where $A^d(s;\theta) = \int_0^s \psi(u;\theta)\alpha(u;\theta)\,\mathrm{d}u$ is the derivative of $A(s;\theta)$ with respect to θ. Before we go on to the mse estimators we note that in numerical integration techniques are called form in most situations when estimating $\widehat{J'}$ and $\widehat{Q'}(t)$.

4.3.1 Cumulative hazard

We start out with the case where estimation of the cumulative hazard function is of interest. As usual we spilt the mse into the two terms of squared bias and variance, and estimate these separately.

Since it is seen from corollary 4.2.4 that the limiting distribution of $\sqrt{n}(\widehat{A}_{\text{np}}(t) - A_{\text{true}}(t))$ has mean zero, it is natural to let

$$\widehat{\text{bias}^2}_{A,\text{np}} = 0.$$

From the same limiting distribution, we see that the natural estimate of the variance of the nonparametric estimator is

$$\widehat{\text{Var}}(\widehat{A}_{\text{np}}(t)) = \frac{1}{n}\widehat{V'}_{A,\text{np}}(t) = \frac{1}{n}\widehat{\nu'}(t).$$

Thus, an estimator for the mean squared error of the nonparametric estimator for the cumulative hazard function is

$$\widehat{\text{mse}}_{A,\text{np}} = \widehat{\text{bias}^2}_{A,\text{np}} + \widehat{\text{Var}}(\widehat{A}_{\text{np}}(t)) = \frac{1}{n}\widehat{V'}_{\text{np}}(t).$$

Regarding the parametrics, the variance is naturally estimated by

$$\widehat{\text{Var}}(\widehat{A}_{\text{pm}}(t)) = \frac{1}{n}\widehat{V'}_{A,\text{pm}}(t).$$

For the squared bias, we use the same estimation strategy as for the regular iid case of chapter 3, i.e. we estimate the parametric bias by

$$\widehat{\text{bias}}_{\text{pm}} = \widehat{b}_A(t) = \widehat{A}_{\text{pm}}(t) - \widehat{A}_{\text{np}}(t).$$

From the third limiting distribution of corollary 4.2.4 it is seen that $\widehat{b}_A(t)$ is an asymptotically unbiased estimator for $b_A(t)$ which also is asymptotically close to the true bias: $E_G[\widehat{A}_{\text{pm}}(t) - A_{\text{true}}(t)]$. We are estimating the squared bias, and as seen earlier in the thesis, just squaring such a quantity will tend to overestimate the squared bias by an additional variance term. We therefore subtract an estimate of the variance of the bias estimator to obtain the following squared bias estimator

$$\begin{aligned}\widehat{\text{bias}^2}_{\text{pm}} &= \widehat{b}_A^2(t) - \frac{1}{n}\widehat{V'}_{b,A}(t) \\ &= (\widehat{A}_{\text{pm}}(t) - \widehat{A}_{\text{np}}(t))^2 - \frac{1}{n}\left(\widehat{V'}_{A,\text{pm}}(t) + \widehat{V'}_{A,\text{np}}(t) - 2\widehat{V'}_{A,\text{pm,np}}(t)\right).\end{aligned}$$

In total we then get the following mse estimator in the parametric case:

$$\begin{aligned}\widehat{\text{mse}}_{A,\text{pm}} &= \widehat{\text{bias}^2}_{A,\text{pm}} + \widehat{\text{Var}}(\widehat{A}_{\text{pm}}(t)) \\ &= (\widehat{A}_{\text{pm}}(t) - \widehat{A}_{\text{np}}(t))^2 - \frac{1}{n}\left(\widehat{V'}_{A,\text{np}}(t) + 2\widehat{V'}_{A,\text{pmnp}}(t)\right).\end{aligned}$$

By collecting the mse estimators in the two situations, we finally define a FIC scheme choosing the model, whose following FIC score is the smallest:

$$\begin{aligned}\text{FIC}_{\text{np}} &= \frac{1}{n}\widehat{V'}_{A,\text{np}}(t), \\ \text{FIC}_{\text{pm}} &= (\widehat{A}_{\text{pm}}(t) - \widehat{A}_{\text{np}}(t))^2 - \frac{1}{n}\left(\widehat{V'}_{A,\text{np}}(t) - 2\widehat{V'}_{A,\text{pm,np}}(t)\right).\end{aligned}$$

4.3.2 Survival probability

Similarly, a FIC scheme may be derived in the case where the survival probability is the focus. The estimators are analogous and we will therefore shorten this section by only giving the estimators without any further explanation:

$$\begin{aligned}
\widehat{\text{bias}^2}_{S,\text{np}} &= 0, \\
\widehat{\text{Var}}(\widehat{S}_{\text{np}}(t)) &= \frac{1}{n}\widehat{V'}_{S,\text{np}}(t), \\
\widehat{\text{bias}^2}_{S,\text{pm}} &= \widehat{b}^2_S(t) - \frac{1}{n}\widehat{V'}_{b,S}(t) \\
&= (\widehat{S}_{\text{pm}}(t) - \widehat{S}_{\text{np}}(t))^2 - \frac{1}{n}\left(\widehat{V'}_{S,\text{pm}}(t) + \widehat{V'}_{S,\text{np}}(t) - 2\widehat{V'}_{S,\text{pm,np}}(t)\right), \\
\widehat{\text{Var}}(\widehat{S}_{\text{pm}}(t)) &= \frac{1}{n}\widehat{V'}_{S,\text{pm}}(t).
\end{aligned}$$

These estimators then gives the following FIC scores:

$$\text{FIC}_{\text{np}} = \frac{1}{n}\widehat{V'}_{S,\text{np}}(t), \tag{4.11}$$

$$\text{FIC}_{\text{pm}} = (\widehat{S}_{\text{pm}}(t) - \widehat{S}_{\text{np}}(t))^2 - \frac{1}{n}\left(\widehat{V'}_{S,\text{np}}(t) - 2\widehat{V'}_{S,\text{pm,np}}(t)\right). \tag{4.12}$$

As usual, the FIC criterion is defined as selecting the model with the smallest FIC score.

4.3.3 Additional notes

As in the regular iid situation with uncensored data, one may get misleading estimates using the above formulae. I.e. the possible problem with negatively estimated squared bias and variance may also occur in these situations. The natural solution is then to only include the estimators if they are positively estimated. Using this strategy, the nonparametric FIC scores remains unchanged, whereas the parametric mse estimators changes to

$$\text{FIC}^*_{\text{pm}} = \left\{(\widehat{A}_{\text{pm}}(t) - \widehat{A}_{\text{np}}(t))^2 - \frac{1}{n}\left[\widehat{V'}_{A,\text{pm}}(t) + \widehat{V'}_{A,\text{np}}(t) - 2\widehat{V'}_{A,\text{pm,np}}(t)\right]^+\right\}^+ + \widehat{V'}_{A,\text{pm}}(t),$$

for the cumulative hazard and

$$\text{FIC}^*_{\text{pm}} = \left\{(\widehat{S}_{\text{pm}}(t) - \widehat{S}_{\text{np}}(t))^2 - \frac{1}{n}\left[\widehat{V'}_{S,\text{np}}(t) + \widehat{V'}_{S,\text{pm}}(t) - 2\widehat{V'}_{S,\text{pm,np}}(t)\right]^+\right\}^+ + \widehat{V'}_{S,\text{pm}}(t),$$

for the survival probability. We also note that it should be quite straight forward to create a multivariate extensions of these FIC schemes. Our focus has been on continuous parametric distributions. The extension to discrete parametric distributions may probably be handled in a similar way. The results of Hjort (1992) must however be validated also for discrete parametric distributions. Note also that when there are ties in the data set, i.e. that two or more observations are equal, some minor adjustments are called for. More on the last topic may be found in Aalen et al. (2008, chapter 3.1.3).

4.4 Discussion of conditions

The results in the latest section were based on a limiting distribution which was proved under a number of assumptions. We assumed that the time point t of interest is contained in the interval $(0, \tau)$. This is a natural condition since for t outside this interval we do not have any data available for estimation purposes. The conditions (4.2) and (4.3) cannot be checked in practical situations where one does not know anything about the censoring mechanism. However, what these conditions really say, is that if more individuals were added to the study, the proportion of the individuals at risk at time point s would stabilize at some nonzero value with probability 1 at all time points s. Condition (4.4) is exclusively a mathematical condition used to make theorem B.2.12 work properly. This condition is also hard to check in practical situations, but it is a rather weak assumption. Furthermore it is assumed that $\alpha(s)/y(s)$ is integrable on $[0, t]$, a condition that is satisfied for almost all reasonable function types of both $\alpha(s)$ and $y(s)$.

Regarding the parametrics it is assumed that θ_0 is unique, which is a rather weak assumption. We also assume that both the true hazard function $\alpha(s)$ and the least false hazard function $\alpha(s; \theta_0)$ under the parametric family are bounded away from zero in the interval $(0, \tau)$. This means that in this interval the hazard function must be strictly positive. This is fortunately the case for all natural distributions one would encounter in these situations. For the true hazard function, this cannot be checked, but for the parametric hazard function it can be checked if this is the case for any of the possible values of θ not on the boundary. The key condition (4.5) is on the same form as for the regular iid situation although the quantities involved, are somewhat different. The following lemma is in style of theorem 3.3.3:

Lemma 4.4.1. *Let the data situation be as in explained earlier in this chapter. Suppose also the following conditions hold:*

(a) $\widehat{\theta}_n$ *is the only root of* $\overline{U}_n$ *for every* n *large enough.*

(b) θ_0 *is an interior point of the parameter space* Θ*.*

(c) The hazard function $\alpha(s; \theta)$ *is three times continuously differentiable in a neighborhood of* θ_0*.*

(d) The three derivatives above are dominated by integrable functions of $s \in (0, \tau)$*, independent of* θ*.*

(e) J' *exists, and is nonsingular.*

Under these conditions, relation (4.5) *of assumption 4.2.1 holds.*

Proof. Note that the conditions on $\alpha(s; \theta)$ makes sure that $U(y; \theta)$ is twice continuously differentiable, and that it is also bounded by an integrable function for θ in a neighborhood of θ_0. The result then follows by the same arguments as in theorem 3.3.3. □

Finally, it is assumed that the cdf of the parametric distribution has a nonzero derivative with respect to the parameter vector. Note that it is not assumed that all indices are nonzero, only that not all of them are zero. For a particular situation it could be checked if there are any θ values that may cause this problem. This condition may also here be omitted if one defines 0 as a random variable with expectation and variance 0.

Remark 3. *The derived schemes of this chapter are likely to holds also for more general situations than censored iid data. Aalen (1978) considered the more general problem of martingales consisting of counting process and intensity functions under the multiplicative intensity model. In the above derivations, we used no more than these properties to handle the nonparametric part of the scheme. Hence, the results concerning the nonparametrics should hold also for other data types that may be written on the same form. This opens for competing risk models, birth and death processes, Markov chains with censoring and other types of multi-state models. Studying these models using parametric models is however in general a bit more involved. The necessary derivations corresponding to the treatment of the parametrics in this chapter are also likely be a bit more messy than what we ended up with by concentration exclusively on censored iid data.*

4.5 Other focus parameters

Even if we have been focusing on the two most important and natural focus parameters in this chapter, there may of course be settings where the aim of the model fitting is to estimate a quantity different from these. When the focus parameter can be represented as a simple function of either the survival function or the cumulative hazard, the limiting distribution and hence the mse estimators are quite easily obtained via the delta method. One such situation, is the rather obvious cdf at some point t, which can be written as $\mu(G) = G(t) = 1 - S(t)$. Even if the survival probability most often of greater interest than the less encouraging probability of "death", there are situations where it may be of interesting. It is of greater interest when the data corresponds to something else than death or occurrence of a certain disease. E.g. when the time until a goal is scored in a sport event, such a focus parameter may be of interest, denoting the probability that the event occurs before time t. Investigating this focus parameter further is very simple since the transforming function is just $S_{\text{cdf}}(x) = 1 - x$, which has derivative -1, which in turn gives

$$\sqrt{n}\begin{pmatrix} \widehat{G}_n(t) - G(t) \\ F(t;\widehat{\theta}_n) - F(t;\theta_0) \end{pmatrix} \xrightarrow{d.} \sqrt{n}\begin{pmatrix} \widehat{S}_{\text{np}}(t) - \widehat{S}_{\text{true}}(t) \\ \widehat{S}_{\text{pm}}(t) - S_{0,\text{pm}}(t) \end{pmatrix},$$

where $\widehat{G}_n(t) = 1 - \widehat{S}_{\text{np}}(t)$. Thus, the limiting results for the survival function hold also for the cdf. Since the FIC formulae consist only of squared quantities, the FIC scheme for this situation is exactly analogous as well. Therefore, the FIC scheme with formulae (4.11) and (4.12) may be applied directly to the situation where the cdf at t is the focus as well.

Focus parameters like the median, other quantiles, the expectation, the variance etc., may surely be of interest also for censored data. The parametric part of the limiting distribution is not really problematic, since that is just a matter of formulating the focus parameter in terms of θ and using the delta method. The nonparametric part is however a bit more troublesome. To deal with model selection for some of these focus parameter, especially the ones that can be formulated as a functional of the cdf or survival function, one may apply the functional delta method in a similar way as in chapter 3. I.e. the functional delta method provides the limiting distribution for functionals $\mu(H)$ for some cdf H. To apply this theory some regularity conditions concerning the smoothness of μ must be fulfilled. However, when the functional is Hadamard differentiable, the theory usually works out well. In style with Andersen et al. (1993,

chapter IV.3.4), we have that

$$\sqrt{n}(\mu(\widehat{G}_n) - \mu(G)) \xrightarrow{L} \mathrm{d}\mu(G) \cdot Z,$$

where Z is the process, whose realizations are distributed as given in equation (4.10), and $\mathrm{d}\mu(G)$ is the functional derivative (see definition B.1.1) of μ at G. This is a somewhat different version of the functional delta method, than the one used in the previous chapter. Here we do not to go via the use of influence functions, but apply the more general functional derivative of μ directly. The reason for this is that since we are working with data which we treat by integrating over processes, not by simply summing over variables. The nice representation of a mean of influence functions evaluated at the data points we get for regular iid data, is thus not present in this situation. Therefore, it becomes somewhat more complicated to work out the joint limiting distribution of the nonparametric and parametric estimators. Also, since we are most interested in the cumulative hazard and the survival function, and censored data is not of main interest in this thesis, we do not go further than this. Marginal limiting distributions for the nonparametric estimator of the quantile function and a few other nice functionals are given in Andersen et al. (1993, chapter IV.3.4) for the interested reader.

4.6 Illustration: The simplest survival model

Here we consider the simplest model selection situation for censored data one would encounter of this type: Nonparametrics vs. the exponential distribution, when the focus parameter is the cumulative hazard rate at time t. This is particularly simple as the exponential model has a constant hazard rate. The cumulative hazard rate based on this distribution is therefore simply given by $t\theta$ which will be estimated by $t\widehat{\theta}_n$, whereas the nonparametric estimator is simply the Nelson-Aalen estimator. For this situation we get that

$$\begin{aligned} \alpha(s;\theta_0) &= \theta_0, & \psi(s;\theta_0) &= \frac{1}{\theta_0}, \\ \dot{\psi}(s;\theta_0) &= -\frac{1}{\theta_0^2}, & \frac{\partial \mu_F(\theta)}{\partial \theta} &= t, \end{aligned}$$

which gives

$$\begin{aligned} J' &= \frac{1}{\theta_0^2}\int_0^\tau y(s)\alpha(s)\,\mathrm{d}s, \\ K' &= \frac{1}{\theta_0^2}\int_0^\tau y(s)\alpha(s)\,\mathrm{d}s + 2\int_0^\tau \iota(s)\,\mathrm{d}s, \\ \nu'(s) &= \int_0^s \frac{\alpha(u)}{y(u)}\,\mathrm{d}u, \\ Q'(t) &= \frac{1}{\theta_0}A_{\text{true}}(t) - \frac{1}{\theta_0}\int_0^\tau y(s)(\alpha(s)-\theta_0)\nu'(\max\{s,t\})\,\mathrm{d}s, \end{aligned}$$

where

$$\iota(s) = \frac{1}{\theta_0}\int_0^s y(u)(\alpha(u)-\theta_0)\,\mathrm{d}u.$$

Thus, we end up with the following estimators for the related to the FIC($\widehat{\mu}_{\rm pm}$):

$$\begin{aligned}\widehat{A}_{\rm pm}(t) &= t\widehat{\theta}_n,\\ \widehat{V'}_{A,\rm pm} &= t^2\frac{\widehat{K'}}{\widehat{J'}^2},\\ \widehat{V'}_{A,\rm pm,np} &= t\frac{\widehat{Q'}(t)}{\widehat{J'}},\end{aligned}$$

where

$$\begin{aligned}\widehat{J'} &= \frac{1}{n\widehat{\theta}_n^2}N(\tau),\\ \widehat{K'} &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{D_i}{\widehat{\theta}_n}-T_i\right),\\ \widehat{Q'}(t) &= \frac{1}{\widehat{\theta}_n}\widehat{A}_{\rm np}(t)-\frac{1}{n\widehat{\theta}_n}\sum_{i=1}^{n}\widehat{\nu'}(\max\{T_i,t\})+\frac{1}{n}\int_0^\tau Y(s)\widehat{\nu'}(\max\{s,t\})ds.\end{aligned}$$

With these formulae one may without too much trouble calculate the FIC scores for a concrete situation with these two as competing models. Note however that even in this simple situations, $\widehat{Q'}(t)$ is easiest to obtain using numerical integration.

Chapter 5

Various related FIC topics

In two previous chapters we handled FIC for two specific types of data. It is however clear that focused model selection between parametric and nonparametric models is of interest also for other types of data.

In this chapter we discuss situations with some relation to the topics already handled, where a FIC scheme still might be helpful to perform model selection. These topics are however not discussed to the fullest. Some of the techniques may be applied directly the way they are presented, while others may need to be processed some more in someone's mind before decisions in practical problems are based on them.

We start out by dealing with density estimation for continuous data in the usual setting of iid data. Furthermore, a FIC scheme in the more complex regression framework is sketched, before we provide FIC formulae in some of the most common situations connected to comparison of two iid data sets. Moreover, FIC formulae are derived under the local misspecification framework used to deal with theoretical aspects of the schemes in chapter 3. Then we discuss FIC schemes based solely on resampling techniques, before we finish off by discussing FIC when the parametric distribution does not fit the framework of chapter 3.

5.1 FIC for density estimation

The density of a continuous probability distribution is the analogue of the probability mass function for a discrete probability distribution. Such a measure might be of interest in many different situations since it quantifies how likely the occurrence of a certain value is. Whenever one is interested in such a quantity, the question on how one might estimate it naturally arises. There is a wide range of accessible estimators for this quantity, and therefore also model selection plays an important role for this situation. Our attention is as usual drawn towards choosing between parametrics and nonparametrics.

When basing estimation on parametrics, the natural strategy is to fit the model parameters using the method of maximum likelihood, and then use the estimate $\widehat{\mu}_{\rm pm} = f(y_0; \widehat{\theta}_n)$. If one does not want to rely on a certain parametric family, it all gets a bit more difficult. As we saw in section 3.3 of chapter 3, density estimation by simply plugging in the ecdf in the functional $\mu(H) = \frac{\partial H(y)}{\partial y}\big|_{y=y_0}$ did not work very well. The main problem was that the estimator will be zero with probability 1. An at the outset quite fruitful approach for nonparametric density

estimation, is to use the representation

$$g(y_0) = \lim_{\epsilon \to 0} \frac{G(y_0 + \epsilon) - G(y_0)}{\epsilon},$$

to estimate the density by

$$\widehat{g}_{\text{np}}(y_0) = \frac{\widehat{G}_n(y_0 + \epsilon) - \widehat{G}_n(y_0)}{\epsilon},$$

or seemingly more robust

$$\widehat{g}_{\text{np}}(y_0) = \frac{\widehat{G}_n(y_0 + \epsilon) - \widehat{G}_n(y_0 - \epsilon)}{2\epsilon},$$

for any small $\epsilon > 0$. Both these estimators are consistent when letting ϵ depend on the sample size such that ϵ decreases in a suitable way as n increases. The problem with such an approach for a finite sample, is however that it depends heavily on the chosen ϵ. For too small samples one must also have to choose a rather big ϵ to get a nonzero estimate. The problem using nonparametrics for this type of estimation problem is connected to the fact that we rely on something unsmooth. Smoothing is exactly the key in nonparametric density estimation. The clearly most applied smoothing techniques are those of the kernel smoothing type. In particular, one then uses estimators on the form

$$\widehat{g}_n(y_0) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{y_0 - Y_i}{h} \right),$$

for a kernel function K and bandwidth h both chosen in advance. The kernel function K must integrate to 1 and should be symmetric. From this point throughout the section, will be working with $\widehat{\mu}_{\text{np}} = \widehat{g}_n(y_0)$ for given K and h. There are however no obvious choices of neither K nor h. Studies have however shown that the Epanechnikov kernel given by

$$K(y) = \frac{1}{c} p\left(\frac{y}{c} \right),$$

where $p(u) = \frac{3}{4}(1 - u^2)\mathbf{1}_{\{|u| \leq 1\}}(u)$, is optimal in terms of mse minimization for the first order large sample approximation of the problem. Apart from this, using the normal distribution has been very popular and it also works out fairly well in most cases. The choice of bandwidth h is a more comprehensive task, and since it also indicates the degree of smoothing, it is obvious that wrong choices may lead to estimators which are clearly off target. For large samples it is natural to smooth less than for small samples, thus the bandwidth h should depend on n and decrease as n increases. Especially the following nice corollary is given in Lehmann (1998):

Corollary 5.1.1. (Consistency of the kernel density estimator, slightly rewritten from Lehmann (1998, corollary 6.4.1))
Let $g(y)$ being three times differentiable with bounded third derivative in a neighborhood of y_0, K be symmetric about 0 with

$$\int K^2(y)\,\mathrm{d}y, \quad \int y^2 K(y)\,\mathrm{d}y \quad \textit{and} \quad \int |y|^3 K(y)\,\mathrm{d}y$$

all being finite. Finally $h = h_n$ depends on n such that

$$h_n \to 0, \quad nh_n \to \infty \text{ as } n \to \infty.$$

Then $\widehat{g}_n(y_0) = \frac{1}{n}\sum_{i=1}^n \frac{1}{h_n} K\left(\frac{y_0 - Y_i}{h_n}\right)$ is a consistent estimator for $g(y_0)$.

Proof. A Taylor expansion is used on $g(y_0)$ to give first order approximations for the bias and variance of $\widehat{g}_n(y_0)$. One then shows that both these terms (including remainders) converge to zero, and the results follows. The complete proof is given in Lehmann (1998, corollary 6.4.1 and theorem 6.4.3). □

More on kernel estimation may be found in e.g. Wasserman (2006, chapter 6.3), or Silverman (1986). For now, let us assume that we have chosen a kernel function K and a bandwidth h without knowing the data which means that these can be treated as nonstochastic variables.

5.1.1 Mean squared error estimators

To follow the FIC idea and choose model based on which model that has the smallest estimated mse, we need estimators of the squared bias and variance. Under the conditions stated in corollary 5.1.1, the bias and variance of the nonparametric density estimator may be written as

$$\begin{aligned}
\text{bias}(\widehat{g}_n(y_0)) &= \frac{1}{2} h_n^2 g^{(2)}(y_0)\tau^2 + o(h_n^2), \\
\text{Var}(\widehat{g}_n(y_0)) &= \frac{1}{nh_n} g(y_0) K_2 + o\left(\frac{1}{nh_n}\right),
\end{aligned}$$

where $g^{(2)}(y_0) = \frac{\partial^2}{\partial y^2} g(y)\big|_{y=y_0}$, $\tau^2 = \int y^2 K(y)\,\mathrm{d}y$ and $K_2 = \int K^2(y)\,\mathrm{d}y$. Note that we use "$^{(2)}$" to denote the second derivative. Since both h_n^2 and $1/(nh_n)$ converges to zero as n grows, we may use the first term in each of these formulae as asymptotic approximations. The formulae are though not directly computable since they contain the factors $g^{(2)}(y_0)$ and $g(y_0)$, which are obviously unknown. The latter may however be estimated by $\widehat{g}_n(y_0)$. The former is somewhat more involved, but a natural approach is to differentiate $\widehat{g}_n(y)$ and evaluate it at $y = y_0$. This method does however require that K is twice differentiable. In return one gets the estimator

$$\widehat{g^{(2)}}_n(y) = \frac{1}{n}\sum_{i=1}^n \frac{1}{h^3} K^{(2)}\left(\frac{y - Y_i}{h}\right).$$

For the estimator to be consistent one might have to add some additional regularity conditions, see Prakasa Rao (1983, Theorem 4.1.1). For a given bandwidth h and kernel function K we arrive at the following estimators for bias and variance of the kernel density estimator:

$$\begin{aligned}
\widehat{\text{bias}}(\widehat{g}_n(y_0)) &= \frac{1}{2} h^2 \widehat{g^{(2)}}_n(y_0)\tau^2, \\
\widehat{\text{Var}}(\widehat{g}_n(y_0)) &= \frac{1}{nh} \widehat{g}_n(y_0) K_2,
\end{aligned}$$

assuming that τ^2 and K_2 can be calculated exactly or approximately by using numerical integration. To estimate the squared bias we should also in this situation adjust the squared bias estimate by subtracting an estimate of its variance. Since $\widehat{g^{(2)}}_n(y_0)$ is the only stochastic term, we get that

$$\text{Var}\left(\text{bias}(\widehat{g}_n(y_0))\right) = \frac{1}{4}h^4\tau^4\text{Var}\left(\widehat{g^{(2)}}_n(y_0)\right).$$

We therefore need an estimate of $\text{Var}(\widehat{g^{(2)}}_n(y_0))$. Since $\widehat{g^{(2)}}_n(y_0)$ is just a mean over transformed iid variables, the variance may however be rewritten as:

$$\text{Var}\left(\widehat{g^{(2)}}_n(y_0)\right) = \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}\left(\frac{1}{h^3}K\left(\frac{y_0-Y_i}{h}\right)\right) = \frac{1}{n}\text{Var}\left(\frac{1}{h^3}K^{(2)}\left(\frac{y_0-Y_i}{h}\right)\right).$$

By using the standard sample variance estimator, we then get

$$\begin{aligned}\widehat{\text{Var}}\left(\widehat{g^{(2)}}_n(y_0)\right) &= \frac{1}{n}\widehat{\text{Var}}\left(\frac{1}{h^3}K^{(2)}\left(\frac{y_0-Y_i}{h}\right)\right)\\ &= \frac{1}{n}\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{1}{h^3}K^{(2)}\left(\frac{y_0-Y_i}{h}\right) - \widehat{g^{(2)}}_n(y_0)\right)^2.\end{aligned}$$

The full estimator for the nonparametric squared bias is therefore given by

$$\begin{aligned}\widehat{\text{bias}^2}(\widehat{\mu}_{\text{np}}) &= \frac{1}{4}h^4\widehat{g^{(2)}}_n(y_0)^2\tau^4 - \frac{1}{4}h^4\tau^4\frac{1}{n}\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{1}{h^3}K^{(2)}\left(\frac{y_0-Y_i}{h}\right) - \widehat{g^{(2)}}_n(y_0)\right)^2\\ &= \frac{1}{4}h^4\tau^4\left(\widehat{g^{(2)}}_n(y_0)^2 - \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{1}{h^3}K^{(2)}\left(\frac{y_0-Y_i}{h}\right) - \widehat{g^{(2)}}_n(y_0)\right)^2\right),\end{aligned}$$

whereas the variance, as earlier seen, may be estimated by

$$\widehat{\text{Var}}(\widehat{\mu}_{\text{np}}) = \frac{1}{nh}\widehat{g}_n(y_0)K_2.$$

Using the above formulae we get an mse estimator and FIC formula for the nonparametric kernel density estimator given by

$$\begin{aligned}\text{FIC}(\widehat{\mu}_{\text{np}}) &= \widehat{\text{bias}^2}(\widehat{\mu}_{\text{np}}) + \widehat{\text{Var}}(\widehat{\mu}_{\text{np}})\\ &= \frac{1}{4}h^4\tau^4\left(\widehat{g^{(2)}}_n(y_0)^2 - \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{1}{h^3}K^{(2)}\left(\frac{y_0-Y_i}{h}\right) - \widehat{g^{(2)}}_n(y_0)\right)^2\right)\\ &\quad + \frac{1}{nh}\widehat{g}_n(y_0)K_2. \qquad (5.1)\end{aligned}$$

Turning to parametrics, the actual estimator is as mentioned fairly standard. Despite this, it is more involved to estimate its mse. We will throughout this section assume that the regularity conditions in assumption 3.1.1 which concerns parametrics, holds. We then have that

$$\widehat{\theta}_n - \theta_0 = J^{-1}\overline{U}_n + o_p\left(\frac{1}{\sqrt{n}}\right),$$

which by a Taylor expansion gives

$$\widehat{\mu}_{\rm pm} - \mu_{0,\rm pm} = f(y_0;\widehat{\theta}_n) - f(y_0;\theta_0) = \left(\frac{\partial}{\partial\theta} f(y_0;\theta)\Big|_{\theta=\theta_0}\right) J^{-1}\overline{U}_n + o_p\left(1/\sqrt{n}\right).$$

Thus, the variance of $\widehat{\mu}_{\rm pm} = f(y_0;\widehat{\theta}_n)$ may be approximated by the usual formula

$$\mathrm{Var}(\widehat{\mu}_{\rm pm}) = \frac{1}{n}\left(\frac{\partial}{\partial\theta} f(y_0;\theta)\Big|_{\theta=\theta_0}\right) J^{-1}KJ^{-1}\left(\frac{\partial}{\partial\theta} f(y_0;\theta)\Big|_{\theta=\theta_0}\right)^t,$$

which is estimated by inserting $\widehat{\theta}_n$ for θ_0. To estimate the bias $E_G[\widehat{\mu}_{\rm pm} - \mu_{\rm true}]$, we go via the nonparametric estimator, as usual. The usual strategy of just inserting the nonparametric estimator for $\mu_{\rm true}$ gives

$$\widehat{\mu}_{\rm pm} - \widehat{\mu}_{\rm np}.$$

However, since $\widehat{\mu}_{\rm np}$ is seen to generally be a biased estimator of $\mu_{\rm true}$, we insert $\widehat{\mu}_{\rm np} - \widehat{\rm bias}(\widehat{\mu}_{\rm np})$ for $\mu_{\rm true}$ instead. This gives

$$\widehat{\rm bias}(\widehat{\mu}_{\rm pm}) = \widehat{\mu}_{\rm pm} - \left(\widehat{\mu}_{\rm np} - \widehat{\rm bias^2}(\widehat{\mu}_{\rm np})\right).$$

Estimating the squared bias is as usual done by squaring this estimate and subtracting the variance of $\widehat{\rm bias}(\widehat{\mu}_{\rm pm})$. This is a bit more troublesome than usual since we have that

$$\begin{aligned}\mathrm{Var}\left(\widehat{\rm bias}(\widehat{\mu}_{\rm pm})\right) &= \mathrm{Var}\left(\widehat{\mu}_{\rm pm}\right) + \mathrm{Var}\left(\widehat{\mu}_{\rm np}\right) + \mathrm{Var}\left(\widehat{\rm bias}(\widehat{\mu}_{\rm np})\right)\\ &\quad - 2\mathrm{Cov}\left(\widehat{\mu}_{\rm pm},\widehat{\mu}_{\rm np}\right) + 2\mathrm{Cov}\left(\widehat{\mu}_{\rm pm},\widehat{\rm bias}(\widehat{\mu}_{\rm np})\right) - 2\mathrm{Cov}\left(\widehat{\mu}_{\rm np},\widehat{\rm bias}(\widehat{\mu}_{\rm np})\right).\end{aligned}$$

The variance terms in the above expression are easily estimated by the formulae already established. To estimate the covariance terms, we make use of the fact that the estimators involved in the covariance terms are all (to a first order of approximation) means of iid variables. We start out with the covariance between the parametric and nonparametric μ estimators, and get that

$$\begin{aligned}\mathrm{Cov}\left(\widehat{\mu}_{\rm pm},\widehat{\mu}_{\rm np}\right) &\approx \mathrm{Cov}\left(\mu_{0,\rm pm} + \left(\frac{\partial}{\partial\theta} f(y_0;\theta)\Big|_{\theta=\theta_0}\right) J^{-1}\overline{U}_n, \frac{1}{n}\sum_{i=1}^n \frac{1}{h}K\left(\frac{y_0 - Y_i}{h}\right)\right)\\ &= \frac{1}{n^2}\left(\frac{\partial}{\partial\theta} f(y_0;\theta)\Big|_{\theta=\theta_0}\right) J^{-1}\mathrm{Cov}\left(\sum_{i=1}^n U(Y_i;\theta_0), \sum_{i=1}^n \frac{1}{h}K\left(\frac{y_0 - Y_i}{h}\right)\right)\\ &= \frac{1}{n^2}\left(\frac{\partial}{\partial\theta} f(y_0;\theta)\Big|_{\theta=\theta_0}\right) J^{-1}\sum_{i=1}^n\sum_{j=1}^n \mathrm{Cov}\left(U(Y_i;\theta_0), \frac{1}{h}K\left(\frac{y_0 - Y_j}{h}\right)\right)\\ &= \frac{1}{n^2}\left(\frac{\partial}{\partial\theta} f(y_0;\theta)\Big|_{\theta=\theta_0}\right) J^{-1}\sum_{i=1}^n E\left[U(Y_i;\theta_0)\frac{1}{h}K\left(\frac{y_0 - Y_i}{h}\right)\right]\\ &= \frac{1}{nh}\left(\frac{\partial}{\partial\theta} f(y_0;\theta)\Big|_{\theta=\theta_0}\right) J^{-1}E\left[U(Y_i;\theta_0)K\left(\frac{y_0 - Y_i}{h}\right)\right], \qquad (5.2)\end{aligned}$$

where we have used the independence of Y_i and Y_j for $i \neq j$ and that $E_G[U(Y_i;\theta_0)] = 0$. The expectation in equation (5.2) may easily be estimated by

$$\frac{1}{n}\sum_{i=1}^{n} U(Y_i;\widehat{\theta}_n)K\left(\frac{y_0 - Y_i}{h}\right).$$

The other factors may in the usual fashion be estimated by plug-in estimators. We thus get the following covariance estimator

$$\widehat{\mathrm{Cov}}\left(\widehat{\mu}_{\mathrm{pm}},\widehat{\mu}_{\mathrm{np}}\right) = \frac{1}{nh}\left(\frac{\partial}{\partial\theta}f(y_0;\theta)\Big|_{\theta=\widehat{\theta}_n}\right)\widehat{J}^{-1}\frac{1}{n}\sum_{i=1}^{n} U(Y_i;\widehat{\theta}_n)K\left(\frac{y_0 - Y_i}{h}\right).$$

The other covariance terms have similar estimators, which are derived by analogous arguments. These derivations are omitted, but the formulae are as follows:

$$\begin{aligned}
\mathrm{Cov}\left(\widehat{\mu}_{\mathrm{pm}},\widehat{\mathrm{bias}}(\widehat{\mu}_{\mathrm{np}})\right) &= \frac{\tau^2}{2nh}\left(\frac{\partial}{\partial\theta}f(y_0;\theta)\Big|_{\theta=\theta_0}\right)J^{-1}E\left[U(Y_i;\theta_0)K^{(2)}\left(\frac{y_0 - Y_i}{h}\right)\right],\\
\mathrm{Cov}\left(\widehat{\mu}_{\mathrm{np}},\widehat{\mathrm{bias}}(\widehat{\mu}_{\mathrm{np}})\right) &= \frac{h^2\tau^2}{2n}\left(\frac{1}{h^4}E\left[K^{(2)}\left(\frac{y_0 - Y_i}{h}\right)K\left(\frac{y_0 - Y_i}{h}\right)\right] - g^{(2)}(y_0)g(y_0)\right).
\end{aligned}$$

These are consequently estimated by

$$\begin{aligned}
\widehat{\mathrm{Cov}}\left(\widehat{\mu}_{\mathrm{pm}},\widehat{\mathrm{bias}}(\widehat{\mu}_{\mathrm{np}})\right) &= \frac{\tau^2}{2nh}\left(\frac{\partial}{\partial\theta}f(y_0;\theta)\Big|_{\theta=\widehat{\theta}_n}\right)\widehat{J}^{-1}\frac{1}{n}\sum_{i=1}^{n} U(Y_i;\widehat{\theta}_n)K^{(2)}\left(\frac{y_0 - Y_i}{h}\right),\\
\widehat{\mathrm{Cov}}\left(\widehat{\mu}_{\mathrm{np}},\widehat{\mathrm{bias}}(\widehat{\mu}_{\mathrm{np}})\right) &= \frac{h^2\tau^2}{2n}\left(\frac{1}{h^4}\frac{1}{n}\sum_{i=1}^{n} K^{(2)}\left(\frac{y_0 - Y_i}{h}\right)K\left(\frac{y_0 - Y_i}{h}\right) - \widehat{g^{(2)}}_n(y_0)\widehat{g}_n(y_0)\right).
\end{aligned}$$

Summing up all variance and covariance estimators we get the following estimate for the variance of the bias estimator:

$$\begin{aligned}
\widehat{\mathrm{Var}}\left(\widehat{\mathrm{bias}}(\widehat{\mu}_{\mathrm{pm}})\right) &= \widehat{\mathrm{Var}}(\widehat{\mu}_{\mathrm{pm}}) + \widehat{\mathrm{Var}}(\widehat{\mu}_{\mathrm{np}}) + \widehat{\mathrm{Var}}\left(\widehat{\mathrm{bias}}(\widehat{\mu}_{\mathrm{np}})\right)\\
&\quad - 2\left(\widehat{\mathrm{Cov}}\left(\widehat{\mu}_{\mathrm{pm}},\widehat{\mu}_{\mathrm{np}}\right) + \widehat{\mathrm{Cov}}\left(\widehat{\mu}_{\mathrm{pm}},\widehat{\mathrm{bias}}(\widehat{\mu}_{\mathrm{np}})\right) - \widehat{\mathrm{Cov}}\left(\widehat{\mu}_{\mathrm{np}},\widehat{\mathrm{bias}}(\widehat{\mu}_{\mathrm{np}})\right)\right),
\end{aligned}$$

which furthermore gives the following estimate for squared bias of the parametric μ estimator:

$$\widehat{\mathrm{bias}^2}(\widehat{\mu}_{\mathrm{pm}}) = \widehat{\mathrm{bias}}(\widehat{\mu}_{\mathrm{pm}})^2 - \widehat{\mathrm{Var}}\left(\widehat{\mathrm{bias}}(\widehat{\mu}_{\mathrm{pm}})\right).$$

Using the formulae above we get an mse estimator and FIC formula for the parametric density estimator given by

$$\begin{aligned}
\mathrm{FIC}(\widehat{\mu}_{\mathrm{pm}}) &= \widehat{\mathrm{bias}^2}(\widehat{\mu}_{\mathrm{pm}}) + \widehat{V}_{\mathrm{pm}}\\
&= \widehat{\mathrm{bias}}(\widehat{\mu}_{\mathrm{pm}})^2 + \widehat{\mathrm{Var}}\left(\widehat{\mu}_{\mathrm{np}}\right) + \widehat{\mathrm{Var}}\left(\widehat{\mathrm{bias}}(\widehat{\mu}_{\mathrm{np}})\right)\\
&\quad - 2\Big[\widehat{\mathrm{Cov}}\left(\widehat{\mu}_{\mathrm{pm}},\widehat{\mu}_{\mathrm{np}}\right) - \widehat{\mathrm{Cov}}\left(\widehat{\mu}_{\mathrm{pm}},\widehat{\mathrm{bias}}(\widehat{\mu}_{\mathrm{np}})\right)\\
&\quad + \widehat{\mathrm{Cov}}\left(\widehat{\mu}_{\mathrm{np}},\widehat{\mathrm{bias}}(\widehat{\mu}_{\mathrm{np}})\right)\Big].
\end{aligned} \tag{5.3}$$

To summarize we have derived approximate formulae for the mean squared error of both a the kernel density estimator and the natural parametric density estimator, and used these to create a FIC scheme. For a particular situation the smallest of the expressions (5.1) and (5.3) determines which model should be used to estimate the density at the point y_0. As usual, in situations with more than one parametric model, one just computes formula (5.3) for each of them to select among several models.

Note that we in contrast to most of the earlier derived schemes did not rely on a joint asymptotic distribution for the parametric and nonparametric μ estimator when deriving the mse estimators. Both of the μ estimators do have a normal limit, but it turns out that the kernel density estimator does not converge in $\sqrt{n}$-rate, but rather the unconventional $n^{2/5}$-rate. The somewhat slower convergence is due to nature an unrestricted density. Even if we know $g(y)$ for all y with $|y - y_0| > \epsilon$, it does not tell us anything about $g(y_0)$; only the values very close to y_0 provide information about $g(y_0)$. Considering this fact, it is not possible to establish a joint distribution for the two estimators in a natural way. Despite this unpleasant behavior, we were able to create a FIC scheme.

5.2 FIC in the regression setting

In a general regression setting, it is assumed that one observes data $(Y_1, x_1), \ldots, (Y_n, x_n)$. Here Y_i is the response variable, which we here for simplicity assume is one-dimensional, and x_i is a p-dimensional covariate vector $x_i = (x_{i1}, \ldots, x_{ip})^t$. As an example the response data might be a person's height, where e.g. sex, weight, shoe size and parent's height may be natural covariates. Here we assume that all covariates are fully specified for each response, i.e. no missing data. Furthermore, one assumes the following relationship between the response and the covariates:

$$Y_i = r(x_i) + \sigma\epsilon_i, \tag{5.4}$$

where ϵ_i is a random error term with mean zero and variance 1, and σ^2 is the variance of the response Y_i which we for simplicity assume is independent of x_i. Moreover, r is the function determining the dependence between x_i and Y_i. The usual form of regression models assumes that the function r is linear, i.e. that for a certain parameter vector β:

$$Y_i = x_i^t\beta + \sigma\epsilon_i.$$

When in addition ϵ_i is assumed to follow a standard normal distribution, the stated model forms the clearly most common form of regression: normal linear regression. A normally distributed error term implies that $Y_i \sim N(x_i^t\beta, \sigma^2)$. For this case, the clever method of least squares which minimizes $\sum_{i=1}^n (Y_i - x_i^t\beta)^2$ is equivalent to maximum likelihood estimation of β and σ. More on this approach may be found in any introductory book of statistics, e.g. Rice (2007).

For quite a few situations the assumption of a normal distribution is not reasonable. When the response only takes positive value close to zero, only integer values or is just binary, the normal assumption is clearly not applicable. For such situations a class of models called generalized linear models (GLM) may be applied. These models assume that the response follows a parametric model belonging to the exponential family. The exponential family includes, among others, the normal, exponential, gamma, Weibull, Bernoulli and Poisson distribution, and is hence a fairly general class. For more on GLM models, see e.g. de Jong and Heller (2008).

We have so far introduced regression models based on parametrics. Nonparametrics methods can however also be applied for this type of models. Going back to the general setup in equation (5.4), it may be natural to estimate the r function using nonparametric methods. In particular, we will here concentrate on models which are so-called linear smoothers. That is models where $\widehat{r}_n(x)$ is an estimator of r on the form

$$\widehat{r}_n(x) = \sum_{i=1}^{n} l_i(x) Y_i,$$

for x a p-dimensional covariate vector, and $l_i(x)$ a smoothing function depending on the covariates. It is in this context common to use smoothing functions based on a kernel function K, as discussed in the previous section. One of the most famous smoothers on this form is the Nadaraya–Watson kernel estimator for $\widehat{r}_n(x)$. For a one-dimensional covariate vector it is given by

$$l_i(x) = \frac{K\left(\frac{x-x_i}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{x-x_j}{h}\right)},$$

where h again is a bandwidth smoothing parameter. More on nonparametric regression models may be found in e.g. Wasserman (2006, chapter 5).

5.2.1 A sketch of a FIC scheme

Since there exists both quite natural parametric and nonparametric regression models, model selection is a theme also for this setting. In the spirit of this thesis we discuss the model selection in terms of a certain focus parameter. However, as we will see, it is a tedious task to derive a FIC apparatus for this setting. Even when concentrating on the possibly simplest focus parameter of this setting, the task is overwhelming. We will therefore be content with giving the rough strategy here, also since regression is not an important part of this thesis.

Let us for simplicity assume that both the response Y_i and the covariate x_i are continuous (e.g. not taking only integer values), the covariate vector is one-dimensional and that the Nadaraya–Watson model represents the nonparametric approach. Furthermore, we focus on the quantity $E_G[Y_i|x]$, i.e. the expected value of the response for a certain covariate. The nonparametric estimator then becomes simply $\widehat{\mu}_{\text{np}} = \widehat{r}_n(x)$ and the parametric estimator based on any linear regression scheme is even simpler given by $\widehat{\mu}_{\text{pm}} = x\widehat{\beta}_n$, where $\widehat{\beta}_n$ is the ML estimator of β under the assumed parametric model. As usual we wish to estimate the squared bias and variance of these estimators. Starting with the nonparametrics Wasserman (2006, Theorem 5.65) states that under rather weak regularity conditions

$$\begin{aligned}
\text{bias}(\widehat{\mu}_{\text{np}}) &= h_n^2 \left(\frac{1}{2} r^{(2)}(x) + \frac{\dot{r}(x)\dot{d}(x)}{d(x)} \right) \int x^2 K(x)\, \mathrm{d}x + o_p(h_n^2), \\
\text{Var}(\widehat{\mu}_{\text{np}}) &= \frac{\sigma^2}{d(x) n h_n} \int K^2(x)\, \mathrm{d}x + o_p\left(1/(nh_n)\right),
\end{aligned}$$

where the two o_p-remainders both disappears as the sample size increases. Here d denotes the density of the covariate values. Omitting the o_p-terms these formulae can be used to estimate the bias and variance of $\widehat{\mu}_{\text{np}}$. By some algebra, one arrives at estimators for the

derivatives $\dot{r}(x)$ and $r^{(2)}(x)$ by differentiating $\widehat{r}_n(x)$. Furthermore, one may apply regular kernel density estimation techniques as discussed in section 5.1, to estimate $d(x)$ and its derivative $\dot{d}(x)$. However, when the covariates are fairly regularly spread out, it may seem fairly natural to assume that the covariate distribution is uniform, or that the density of the covariates is constant at x. For both of these cases $\dot{d}(x) = 0$, and the fraction involving $\dot{d}(x)$ disappears. As repeated a number of times, squaring the bias estimator is not enough to estimate the squared bias – an estimator of its variance has to be subtracted as well. Even if we have made some simplifying assumptions, this is no easy task. The estimation trouble are caused by the estimators for $\dot{r}(x)$ and $r^{(2)}(x)$ which have sums of iid variables both in the numerator and denominator. Empirical plug-in estimators are therefore not directly applicable. One may however try to estimate these using resampling techniques as the bootstrap or the jackknife. For an introduction to these estimation techniques, see e.g. Efron and Tibshirani (1993).

Assume for now that we managed to estimate the mse of the nonparametric estimator. Turning to parametrics, one arrives at a normal limiting distribution for $\widehat{\beta}_n$ based on ML estimation in the general GLM setting under fairly weak regularity assumptions. The limiting distribution has the usual form

$$\sqrt{n}(\widehat{\beta} - \beta_0) \xrightarrow{L} N\left(0, J^{-1}KJ^{-1}\right).$$

Since the parametric estimator is given by $x\widehat{\beta}_n$, the asymptotic variance of $\widehat{\mu}_{\text{pm}}$ can be written as

$$\text{Var}(\widehat{\mu}_{\text{pm}}) = x^2 J^{-1}KJ^{-1}.$$

This variance is however easily estimated via plug-in estimators for J and K. To estimate the bias it is as usual natural to go via the nonparametric estimator. As in section 5.1.1, the nonparametric estimator is biased, so an additional bias term for the nonparametric estimator must be included as well. This leads to an estimator on the form $\widehat{\mu}_{\text{pm}} - (\widehat{\mu}_{\text{np}} - \widehat{\text{bias}}(\widehat{\mu}_{\text{np}}))$. To estimate the squared bias, the variance of this estimator is needed. This includes both variance terms and covariance terms, where the former can be established by using estimators already sketch, and the latter is on a form similar to the one we recommended resampling techniques for above. However, assuming that we were able to complete all the estimation tasks we have indicated so far in this subsection, we have all estimates we need to establish a FIC scheme for model selection among these two models.

One could certainly think of many useful expansion of a FIC scheme as indicated above. The nonparametric estimator can easily be extended to handle more than one single covariate by introducing a higher dimensional kernel function K and bandwidth vector h. Of course this also applies to the parametric estimators, since using a vector of covariates is straight forward as noted in the introduction to regression above. One might also imagine that a model selection scheme of this type that compares models including different covariates, would be fruitful. If a covariate z does not add more information to the response, one might be better off leaving it out. A scheme selecting between GLM models with different covariates included and nonparametric models of the Nadaraya–Watson type where also different covariates types are included, may be established by inserting e.g. the biggest nonparametric model (including its estimated bias) as the true model, and otherwise estimate all of the parametric models as indicated above. Introducing a covariate dependent variance σ may also be helpful sometimes, but as noted the models then starts to get quite complicated.

As seen from the sketch in this section it is a rather complicated task to establish a FIC scheme in this situation. An additional problem with this approach is that since there are so

many different quantities to estimate, the uncertainty in the estimation of the quantities may also be quite large. Slight changes in the data may therefore change the ranking of the models, especially if the sample size is small. This is clearly a bigger problem when the FIC formulae are complicated compared to the simpler schemes of the previous chapters. Even if we did not succeed in giving precise and directly applicable formulae for this setting, the section showed several difficulties regarding the estimation process. This certainly indicates that our approach is harder to transform and fully work out in settings more complex than that of iid data.

5.3 FIC for comparing two samples

The FIC schemes we have considered so far may be used in different situations where there is a single sample only. In some situations one may be interested in a focus parameter that depends on several samples. The classical example of such a situation is the difference between the means of two samples. Such a quantity may clearly be estimated both based on parametrics and nonparametrics. As a consequence, model selection techniques should be used to select which model one should base further inference on. In this section we discuss information criteria to select between models in this situation. We will investigate focused model selections for two of the most natural types of focus parameters comparing two samples. Firstly we consider focus parameters which is a difference of two focus parameters, and then we consider a focus parameter which is a product of two focus parameters. We will also restrict the research to the iid situation considered in chapter 3.

In mathematical terms we assume the following situation: $Y_1, \ldots, Y_n$ are iid random variable with a common distribution whose cdf is given by G_1, and denoted "the first sample". Similarly $X_1, \ldots, X_n$ are iid random variable with a common distribution whose cdf is given by G_2, and is denoted "the second sample". The two samples are also assumed to be independent of each other. In addition we assume that both samples and any focus parameters defined on the samples satisfy the regularity conditions in assumption 3.1.1.

5.3.1 Difference of two focus parameters

Consider now the situation where the focus parameter is the difference of two individual focus parameters where each of them depends only on one of the samples, and not both on the same sample. In mathematical terms we write this as a functional of the following form

$$\mu = \mu(G_1, G_2) = \mu_1(G_1) - \mu_2(G_2), \tag{5.5}$$

where μ_1 and μ_2 are functionals defined only on respectively the first and second sample. For this type of focus parameters, we consider the following type of estimators:

$$\widehat{\mu} = \mu(\widehat{G}_1, \widehat{G}_2) = \mu_1(\widehat{G}_1) - \mu_2(\widehat{G}_2),$$

for some estimators $\widehat{G}_1, \widehat{G}_2$ of the cdfs G_1, G_2. Still working in the "parametric vs. nonparametric" world, these estimates will typically consist of $\widehat{G}_n$ and $F_{\widehat{\theta}_n}$ for the two data sets. As usual we use the mse as a measure of the uncertainty of an estimator like in equation (5.5). Note that since the samples are independent the covariance between $\mu_1(\widehat{G}_1)$ and $\mu_2(\widehat{G}_2)$ is zero. For

our convenience, let us write Var_{G_i} for $\text{Var}_{G_i}(\mu_i(\widehat{G}_i))$ and bias_{G_i} for $\text{bias}_{G_i}(\mu_2(\widehat{G}_i))$ for $i = 1, 2$ in addition to Cov_{G_1,G_2} for $\text{Cov}_{G_1,G_2}(\mu_1(\widehat{G}_1), \mu_2(\widehat{G}_2))$. We then get

$$\begin{aligned}
\text{mse}(\mu(\widehat{G}_1, \widehat{G}_2)) &= \text{bias}_{G_1,G_2}(\mu(\widehat{G}_1, \widehat{G}_2))^2 + \text{Var}_{G_1,G_2}(\mu(\widehat{G}_1, \widehat{G}_2)) \\
&= (\text{bias}_{G_1} - \text{bias}_{G_2})^2 + \text{Var}_{G_1} + \text{Var}_{G_2} - 2\text{Cov}_{G_1,G_2} \\
&= \text{bias}_{G_1}^2 + \text{bias}_{G_2}^2 - 2\text{bias}_{G_1}\text{bias}_{G_2} + \text{Var}_{G_1} + \text{Var}_{G_2} \\
&= \text{mse}(\mu_1(\widehat{G}_1)) + \text{mse}(\mu_2(\widehat{G}_2)) - 2\text{bias}_{G_1}\text{bias}_{G_2},
\end{aligned}$$

We do actually get a mse-formula that adds the two marginal mean squared errors, and subtracts a correction term. The correction term reduces the error in the case where the bias of the estimators has the same sign and increases the error when they have different signs.

To estimate this quantity we need estimators of the unsquared biases for the estimators of both μ_1 and μ_2. This is however directly provided by $\widehat{\mu}_1 - \widehat{\mu}_{1,\text{np}}$ and $\widehat{\mu}_2 - \widehat{\mu}_{2,\text{np}}$ since $\widehat{\mu}_{1,\text{np}}$ and $\widehat{\mu}_{2,\text{np}}$ are unbiased estimators under the usual conditions. Hence, the natural estimator for this mse is given by

$$\widehat{\text{mse}}(\mu(\widehat{G}_1, \widehat{G}_2)) = \widehat{\text{mse}}(\mu_1(\widehat{G}_1)) + \widehat{\text{mse}}(\mu_2(\widehat{G}_2)) - 2(\widehat{\mu}_1 - \widehat{\mu}_{1,\text{np}})(\widehat{\mu}_2 - \widehat{\mu}_{2,\text{np}}). \tag{5.6}$$

For the simplest case of just one parametric model, which are used for both samples, we get the following estimators:

1. Nonparametric + nonparametric: $\widehat{\mu}_{\text{np,np}} = \widehat{\mu}_{1,\text{np}} - \widehat{\mu}_{2,\text{np}}$.
2. Nonparametric + parametric: $\widehat{\mu}_{\text{pm,np}} = \widehat{\mu}_{1,\text{pm}} - \widehat{\mu}_{2,\text{np}}$.
3. Parametric + nonparametric: $\widehat{\mu}_{\text{np,pm}} = \widehat{\mu}_{1,\text{np}} - \widehat{\mu}_{2,\text{pm}}$.
4. Parametric + parametric: $\widehat{\mu}_{\text{pm,pm}} = \widehat{\mu}_{1,\text{pm}} - \widehat{\mu}_{2,\text{pm}}$.

For these estimators, equation (5.6) motivates the following mse estimators

$$\begin{aligned}
\text{FIC}(\widehat{\mu}_{\text{np,np}}) &= \frac{1}{n}\widehat{V}_{1,\text{np}} + \frac{1}{m}\widehat{V}_{2,\text{np}}, \\
\text{FIC}(\widehat{\mu}_{\text{pm,np}}) &= (\widehat{\mu}_{1,\text{pm}} - \widehat{\mu}_{1,\text{np}})^2 - \frac{1}{n}\widehat{V}_{1,\text{np}} + 2\frac{1}{n}\widehat{V}_{1,\text{pm,np}} + \frac{1}{m}\widehat{V}_{2,\text{np}}, \\
\text{FIC}(\widehat{\mu}_{\text{np,pm}}) &= \frac{1}{n}\widehat{V}_{1,\text{np}} + (\widehat{\mu}_{2,\text{pm}} - \widehat{\mu}_{2,\text{np}})^2 - \frac{1}{m}\widehat{V}_{2,\text{np}} + 2\frac{1}{m}\widehat{V}_{2,\text{pm,np}}, \\
\text{FIC}(\widehat{\mu}_{\text{pm,pm}}) &= (\widehat{\mu}_{1,\text{pm}} - \widehat{\mu}_{1,\text{np}})^2 - \frac{1}{n}\widehat{V}_{1,\text{np}} + 2\frac{1}{n}\widehat{V}_{1,\text{pm,np}} + (\widehat{\mu}_{2,\text{pm}} - \widehat{\mu}_{2,\text{np}})^2 - \frac{1}{m}\widehat{V}_{2,\text{np}} \\
&\quad + 2\frac{1}{m}\widehat{V}_{2,\text{pm,np}} - 2(\widehat{\mu}_{1,\text{pm}} - \widehat{\mu}_{1,\text{np}})(\widehat{\mu}_{2,\text{pm}} - \widehat{\mu}_{2,\text{np}}).
\end{aligned}$$

Note that the correction term is nonzero only in the last estimator consisting of only parametric estimators. As usual, the FIC scheme chooses the estimator with the smallest FIC value.

5.3.2 Product of two focus parameters

The form of the focus parameter in the above section is maybe the most useful one. However, in some cases one might want to take a look at focus parameters on a slightly different form. Consider a focus parameter on the following multiplicative form:

$$\mu = \mu(G_1, G_2) = \mu_1(G_1)\mu_2(G_2),$$

where μ_1 and μ_2 are functionals defined on respectively the first and second sample. Like in the previous section we derive the mse of this focus parameter for a general estimator where $\widehat{G}_1$ and $\widehat{G}_2$ are inserted to estimate respectively G_1 and G_2. When denoting the expectation of the estimators $\mu_1(\widehat{G}_1)$ and $\mu_2(\widehat{G}_2)$ by respectively E_{G_1} and E_{G_2}, and otherwise using the notation of the previous section, we get

$$\begin{aligned}
\text{mse}(\mu(\widehat{G}_1, \widehat{G}_2)) &= \text{bias}_{G_1,G_2}(\mu(\widehat{G}_1, \widehat{G}_2))^2 + \text{Var}_{G_1,G_2}(\mu(\widehat{G}_1, \widehat{G}_2)) \\
&= (E_{G_1}E_{G_2} - \mu_{1,\text{true}}\mu_{2,\text{true}})^2 + \mu_{1,\text{true}}^2\text{Var}_{G_2} + \mu_{2,\text{true}}^2\text{Var}_{G_1} \\
&\quad + \text{Var}_{G_2}\text{Var}_{G_2} \\
&= E_{G_1}^2E_{G_2}^2 + \mu_{1,\text{true}}^2\mu_{2,\text{true}}^2 - 2E_{G_1}E_{G_2}\mu_{1,\text{true}}\mu_{2,\text{true}} \\
&\quad + \mu_{1,\text{true}}^2\text{Var}_{G_2} + \mu_{2,\text{true}}^2\text{Var}_{G_1} + \text{Var}_{G_2}\text{Var}_{G_2} \\
&= \text{mse}(\mu_1(\widehat{G}_1))\text{mse}(\mu_2(\widehat{G}_2)) + E_{G_1}^2E_{G_2}^2 \\
&\quad - 2E_{G_1}E_{G_2}\mu_{1,\text{true}}\mu_{2,\text{true}}.
\end{aligned}$$

Using the same estimators as earlier, we get the following natural mse estimator:

$$\begin{aligned}
\widehat{\text{mse}}(\mu(\widehat{G}_1, \widehat{G}_2)) &= \widehat{\text{mse}}(\mu_1(\widehat{G}_1))\widehat{\text{mse}}(\mu_2(\widehat{G}_2)) + \mu_1(\widehat{G}_1)^2\mu_2(\widehat{G}_2)^2 \\
&\quad - 2\mu_1(\widehat{G}_1)\mu_2(\widehat{G}_2)\widehat{\mu}_{1,\text{np}}\widehat{\mu}_{2,\text{np}}. \qquad (5.7)
\end{aligned}$$

For the simplest case of just one parametric model, which we apply to both samples, we get four natural estimators for μ:

1. Nonparametric + nonparametric: $\widehat{\mu}_{\text{np,np}} = \widehat{\mu}_{1,\text{np}}\widehat{\mu}_{2,\text{np}}$.
2. Nonparametric + parametric: $\widehat{\mu}_{\text{pm,np}} = \widehat{\mu}_{1,\text{pm}}\widehat{\mu}_{2,\text{np}}$.
3. Parametric + nonparametric: $\widehat{\mu}_{\text{np,pm}} = \widehat{\mu}_{1,\text{np}}\widehat{\mu}_{2,\text{pm}}$.
4. Parametric + parametric: $\widehat{\mu}_{\text{pm,pm}} = \widehat{\mu}_{1,\text{pm}}\widehat{\mu}_{2,\text{pm}}$.

For these estimators, equation (5.7) motivates the following mse estimators:

$$\begin{aligned}
\text{FIC}(\widehat{\mu}_{\text{np,np}}) &= \frac{1}{nm}\widehat{V}_{1,\text{np}}\widehat{V}_{2,\text{np}} - \widehat{\mu}_{1,\text{np}}^2\widehat{\mu}_{2,\text{np}}^2, \\
\text{FIC}(\widehat{\mu}_{\text{pm,np}}) &= \left((\widehat{\mu}_{1,\text{pm}} - \widehat{\mu}_{1,\text{np}})^2 - \frac{1}{n}\widehat{V}_{1,\text{np}} + 2\frac{1}{n}\widehat{V}_{1,\text{pm,np}}\right)\frac{1}{m}\widehat{V}_{2,\text{np}} + \widehat{\mu}_{1,\text{pm}}^2\widehat{\mu}_{2,\text{np}}^2 \\
&\quad - 2\widehat{\mu}_{1,\text{pm}}\widehat{\mu}_{1,\text{np}}\widehat{\mu}_{2,\text{np}}^2, \\
\text{FIC}(\widehat{\mu}_{\text{np,pm}}) &= \frac{1}{n}\widehat{V}_{1,\text{np}}\left((\widehat{\mu}_{2,\text{pm}} - \widehat{\mu}_{2,\text{np}})^2 - \frac{1}{m}\widehat{V}_{2,\text{np}} + 2\frac{1}{m}\widehat{V}_{2,\text{pm,np}}\right) + \widehat{\mu}_{1,\text{np}}^2\widehat{\mu}_{2,\text{pm}}^2 \\
&\quad - 2\widehat{\mu}_{1,\text{np}}^2\widehat{\mu}_{2,\text{pm}}\widehat{\mu}_{2,\text{np}}, \\
\text{FIC}(\widehat{\mu}_{\text{pm,pm}}) &= \left((\widehat{\mu}_{1,\text{pm}} - \widehat{\mu}_{1,\text{np}})^2 - \frac{1}{n}\widehat{V}_{1,\text{np}} + 2\frac{1}{n}\widehat{V}_{1,\text{pm,np}}\right)\left((\widehat{\mu}_{2,\text{pm}} - \widehat{\mu}_{2,\text{np}})^2\right. \\
&\quad \left. - \frac{1}{m}\widehat{V}_{2,\text{np}} + 2\frac{1}{m}\widehat{V}_{2,\text{pm,np}}\right) + \widehat{\mu}_{1,\text{pm}}^2\widehat{\mu}_{2,\text{pm}}^2 - 2\widehat{\mu}_{1,\text{pm}}\widehat{\mu}_{2,\text{pm}}\widehat{\mu}_{1,\text{np}}\widehat{\mu}_{2,\text{np}}.
\end{aligned}$$

Also here, the FIC scheme chooses the μ estimators with the smallest FIC value.

5.3.3 Generalizations

The formulae above with only two different samples may be generalized to three or more samples. The formulae and estimators then become much more complicated and are therefore omitted. There may also be situations where a type of comparison different from the two stated types, is of interest. In such situations the most natural approach is to write out the estimator in terms of quantities that can be estimated by one of the samples. Precise mse formulae may then be carried out by carefully rewriting each the squared bias and variance in terms of quantities that can be estimated from the data.

5.4 FIC in the local misspecification framework

In section 3.4.4 and 3.5.2 a local misspecification framework was used to deal with certain theoretical aspects of the main FIC formulae for iid data. In appendix A the results applied in those sections are derived. The key result of this appendix is the joint limiting distribution for the nonparametric and parametric estimators under that particular framework. It would then be appealing, not only to use the framework for theoretical justification of a derived FIC scheme, but to assume such a framework from the start and use it to build a new FIC scheme. This section is devoted to exactly this task. We give precise limiting formulae for the mean squared error of the estimators and furthermore use empirical approximations to the quantities involved to create mse estimators and hence FIC formulae. We will mainly focus on the task of simply one parametric model, but towards the end we will indicate how the idea may be expanded to several parametric models.

Throughout this section we will assume that we got iid data $Y_1, \ldots, Y_n$[1] stemming for the distribution with density on the form

$$g_n(y) = f(y;\theta_0) + \frac{r(y)}{\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right),$$

where $r(y) : \mathbb{R} \to \mathbb{R}$ is assumed to be a function independent on the sample size n, not necessarily continuous, but with the property that $\int r(y)\,\mathrm{d}\upsilon(y) = 0$. As usual $f(y;\theta)$ is the density or pmf of a parametric distribution. Consequently, the cdf of this distribution may be written as

$$G_n(y) = F(y;\theta_0) + \frac{R(y)}{\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right),$$

for $R(y) = \int_{-\infty}^{y} r(x)\,\mathrm{d}\upsilon(y)$, when an additional condition regarding integrability of the o-term is assumed. Such details can be found in appendix A. Furthermore, we assume that the ingredients of assumption A.0.1 holds, such that the result of lemma A.0.2 is valid. The mentioned lemma makes sure that the following limiting distribution holds

$$\sqrt{n}\begin{pmatrix}\widehat{\mu}_{\mathrm{np}} - \mu_{\mathrm{true}} \\ \widehat{\mu}_{\mathrm{pm}} - \mu_{\mathrm{true}}\end{pmatrix} \xrightarrow{L} \Lambda^* \overset{d.}{=} N_2\left(\begin{pmatrix}\left(\frac{\partial \mu_F}{\partial \theta}\Big|_{\theta_0}\right)^t \\ 0 \\ (K^*)^{-1}q_1 - q_2\end{pmatrix}, \begin{pmatrix}V^*_{\mathrm{np}} & V^*_{\mathrm{pm,np}} \\ V^*_{\mathrm{pm,np}} & V^*_{\mathrm{pm}}\end{pmatrix}\right), \quad (5.8)$$

[1] To be precise Y_i depends also on n and should be represented by Y_{in}, but for comparison with the other situations, we omit this additional n.

where

$$
\begin{aligned}
q_1 &= \int U(y;\theta_0) r(y)\,\mathrm{d}\upsilon(y),\\
q_2 &= \int \mathrm{IF}_\mu(y;F_{\theta_0}) r(y)\,\mathrm{d}\upsilon(y),\\
K^* &= E_{F_{\theta_0}}[U(Y_i;\theta_0)U(Y_i;\theta_0)^t],\\
\nu^* &= \lim_{n\to\infty}\nu_n^* = \lim_{n\to\infty}\mathrm{Var}_{F_{\theta_0}}(\mathrm{IF}_\mu(Y_i;G_n)),\\
Q^* &= \lim_{n\to\infty} Q_n^* = \lim_{n\to\infty}\mathrm{Cov}_{F_{\theta_0}}(U(Y_i;\theta_0),\mathrm{IF}_\mu(Y_i;G_n)),\\
V_{\mathrm{pm}}^* &= \left(\left.\frac{\partial\mu_F}{\partial\theta}\right|_{\theta_0}\right)^t (K^*)^{-1}\left(\left.\frac{\partial\mu_F}{\partial\theta}\right|_{\theta_0}\right),\\
V_{\mathrm{np}}^* &= \nu^*,\\
V_{\mathrm{pm,np}}^* &= \left(\left.\frac{\partial\mu_F}{\partial\theta}\right|_{\theta_0}\right)^t (K^*)^{-1} Q^*.
\end{aligned}
$$

All FIC schemes we have considered so far in this thesis, has been based on estimators of the mse: $\mathrm{mse}(\widehat{\mu}) = E_G[(\widehat{\mu}-\mu_{\mathrm{true}})^2]$. Nevertheless, it is more convenient to consider a slightly different quantity in this situation, namely mse^* defined by

$$
\begin{aligned}
\mathrm{mse}^*(\widehat{\mu}) &= \lim_{n\to\infty} E\left[\left(\sqrt{n}(\widehat{\mu}-\mu_{\mathrm{true}})\right)^2\right]\\
&= \lim_{n\to\infty}\left(E\left[\sqrt{n}(\widehat{\mu}-\mu_{\mathrm{true}})\right]\right)^2 + \mathrm{Var}\left(\sqrt{n}(\widehat{\mu}-\mu_{\mathrm{true}})\right).
\end{aligned}
$$

The reason for using this measure instead in this situation, is that the exact expression can be derived directly from relation (5.8). For the nonparametric and parametric estimators we actually get

$$
\begin{aligned}
\mathrm{mse}^*(\widehat{\mu}_{\mathrm{np}}) &= V_{\mathrm{np}}^*,\\
\mathrm{mse}^*(\widehat{\mu}_{\mathrm{pm}}) &= \left(\left(\left.\frac{\partial\mu_F}{\partial\theta}\right|_{\theta_0}\right)^t (K^*)^{-1} q_1 - q_2\right)^2 + V_{\mathrm{pm}}^*.
\end{aligned}
$$

To create FIC formulae based on these expressions, we have to estimate the quantities involved. Especially we will insert $\widehat{G_n}$ for G_n and $\widehat{\theta}_n$ for θ_0. Furthermore, we will in this finite sample experiment estimate $R(y)$ by $\widehat{R}(y) = \sqrt{n}(\widehat{G}_n(y) - F(y;\widehat{\theta}_n))$ and consequently also $r(y)d\upsilon(y)$ by $d\widehat{R}(y)$. Even if $R(y)$ is actually independent of n, this estimator makes sense since n is simple a given scalar for the finite sample experiment. We also ignore limits and aim at estimating ν_n^* and Q_n^*, not ν^* and Q^* directly. Inserting these estimators gives the following FIC formulae:

$$
\mathrm{FIC}(\widehat{\mu}_{\mathrm{np}}) = \widehat{V}_{\mathrm{np}}^*, \tag{5.9}
$$

$$
\mathrm{FIC}(\widehat{\mu}_{\mathrm{pm}}) = (\widehat{q}_2)^2 - \widehat{V}_{q_2}^* + \widehat{V}_{\mathrm{pm}}^*, \tag{5.10}
$$

where

$$
\widehat{V}_{\mathrm{np}}^* = \int \mathrm{IF}_\mu(y;\widehat{G}_n)^2\,\mathrm{d}F_{\widehat{\theta}_n}(y) - \left(\int \mathrm{IF}_\mu(y;\widehat{G}_n)\,\mathrm{d}F_{\widehat{\theta}_n}(y)\right)^2.
$$

Furthermore

$$\widehat{q}_2 = \int \mathrm{IF}_\mu(y; F_{\widehat{\theta}_n})\,\mathrm{d}\widehat{R}(y) = \sqrt{n}\int \mathrm{IF}_\mu(y; F_{\widehat{\theta}_n})\,\mathrm{d}(\widehat{G}_n(y) - F_{\widehat{\theta}_n}(y)) = \sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\mathrm{IF}_\mu(Y_i; F_{\widehat{\theta}_n}),$$

where we have used that

$$\int \mathrm{IF}_\mu(y; F_{\widehat{\theta}_n})\,\mathrm{d}F_{\widehat{\theta}_n}(y) = 0,$$

and

$$\widehat{V}^*_{q_2} = \frac{1}{n}\sum_{i=1}^{n}\left(\mathrm{IF}_\mu(Y_i; F_{\widehat{\theta}_n}) - \overline{\mathrm{IF}}_{\mu,n}(F_{\widehat{\theta}_n})\right)^2.$$

Finally

$$\widehat{V}^*_{\mathrm{pm}} = \left(\frac{\partial \mu_F}{\partial \theta}\bigg|_{\widehat{\theta}_n}\right)^t (\widehat{K}^*)^{-1}\left(\frac{\partial \mu_F}{\partial \theta}\bigg|_{\widehat{\theta}_n}\right),$$

where

$$\widehat{K}^* = \int U(Y_i;\widehat{\theta}_n)U(Y_i;\widehat{\theta}_n)^t\,\mathrm{d}F_{\widehat{\theta}_n}(y).$$

In formula (5.10), it may seem like the term representing q_1 is forgotten. This is however not the case. The natural estimator of q_1 is

$$\begin{aligned}\widehat{q}_1 &= \int U(y;\widehat{\theta}_n)\,\mathrm{d}\widehat{R}(y) = \sqrt{n}\int U(y;\widehat{\theta}_n)\,\mathrm{d}(\widehat{G}_n(y) - F_{\widehat{\theta}_n}(y))\\ &= \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}U(Y_i;\widehat{\theta}_n) - \int U(y;\widehat{\theta}_n)f(y;\widehat{\theta}_n)\,\mathrm{d}\upsilon(y)\right) = \sqrt{n}\left(0 - \int \frac{\partial}{\partial\theta}f(y;\theta)\Big|_{\theta=\widehat{\theta}_n}\mathrm{d}\upsilon(y)\right)\\ &= -\sqrt{n}\frac{\partial}{\partial\theta}\int f(y;\theta)\,\mathrm{d}\upsilon(y)\Big|_{\theta=\widehat{\theta}_n} = 0,\end{aligned}$$

where the last equality follows provided that derivation and integration can be interchanged. We have also used that the sum of the score functions evaluated at the ML estimator is zero since it is well known that the ML estimator is a root of this equation. Thus, the straight forward natural estimator of q_1 is zero, and it is therefore natural to estimate the whole first term of the squared bias by zero.

This scheme is actually very similar to the main scheme for iid data if we divide the formulae (5.9) and (5.10) by $\sqrt{n}$. In particular, if we choose slightly different estimators, relying more on empirical analogous, we end up with exactly the same scheme. To obtain this results the following changes to estimation has to be carried out: Insert $\mathrm{d}\widehat{G}_n$ everywhere $\mathrm{d}F_{\theta_0}$ is represented instead of $\mathrm{d}F_{\widehat{\theta}_n}$. Doing this simplifies the estimators as well since all integrals reduce to sums instead. Also, a totally wrong parametric model will neither affect estimation of these integrals. Furthermore, instead of trying to estimate the bias of the parametric estimator directly, we note that according to corollary A.0.3, this is the limiting expectation of $\sqrt{n}\widehat{b} = \sqrt{n}(\widehat{\mu}_{\mathrm{pm}} - \widehat{\mu}_{\mathrm{np}})$. The right hand side of this equation is known from data, and it is therefore

natural to use this as an estimator for the parametric bias. The variance correction should then not be $\widehat{V}_{q2}^*$, but rather an estimate of the variance of $\sqrt{n}\widehat{b}$. Using the empirical analogue of the asymptotic variance V_b^*, gives the final estimator. Using $\widehat{G}_n$ instead of $F_{\widehat{\theta}_n}$ can also be seen as a finite sample correction. As $n \to \infty$, the two cdfs should coincide, but for the case of a finite n (which always is the case) $\widehat{G}_n$ will represent the data better than $F_{\widehat{\theta}_n}$. Where F_{θ_0} is represented it is mostly as the limit of G_n. Inserting an estimator for G_n instead of F_{θ_0} is therefore reasonable. With the estimation strategy outlined above we get

$$\begin{aligned}
\text{FIC}(\widehat{\mu}_{\text{np}}) &= \widehat{V}_{\text{np}}, \\
\text{FIC}(\widehat{\mu}_{\text{pm}}) &= \big(\sqrt{n}(\widehat{\mu}_{\text{pm}} - \widehat{\mu}_{\text{np}})\big)^2 - \widehat{V}_b + \widehat{V}_{\text{pm}} \\
&= \big(\sqrt{n}(\widehat{\mu}_{\text{pm}} - \widehat{\mu}_{\text{np}})\big)^2 - \widehat{V}_{\text{np}} + 2\widehat{V}_{\text{pm,np}},
\end{aligned}$$

which is exactly the same formulae as for the main scheme for iid data in chapter 3.

For the situation where there are several parametric models, the derivation of a FIC scheme like this, becomes somewhat more involved. It is quite directly seen that if a parametric model is not a special case of the limiting true distribution with density or pmf $f(y;\theta_0)$, the mse* of the estimator based on this model, will in most cases not be finite. Letting $\widehat{\mu}'_{\text{pm}}$ denote the estimator under such a parametric model, this unfortunate behavior appears whenever $\widehat{\mu}'_{\text{pm}}$ does not coincide with $\widehat{\mu}_{\text{pm}}$ in the limit. Nevertheless, when the additional parametric distribution in fact is a special case of the limiting distribution with density or pmf $f(y;\theta_0)$, the approach may be fruitful. Arguments similar to those used in lemma A.0.2, borrowing techniques from Claeskens and Hjort (2008, chapter 5.2), may be used to derive such a scheme. Even if such a strategy may seem fruitful, the results would highly likely be similar to those derived under the usual iid situation. If one really wants to apply the smoothness of local asymptotic frameworks to choose between parametric models which are related to one another, without deriving a more general scheme, there is a way out. One could, rather heuristically, first apply the FIC scheme selecting between a set of parametric models due to Claeskens and Hjort (2003) to choose the best parametric model, and then use the scheme of this section to select between the nonparametric model and the best parametric model from the first selection stage. We do however not recommend such a technique since one changes the assumed underlying distribution of the data during the model selection step, which clearly is not unproblematic.

5.5 FIC based on resampling

The FIC schemes proposed so far in this thesis are all motivated by large sample results which are used to estimate the mean squared error. The strategy works well in most cases since the approximation errors using asymptotics in finite n situations are most often negligible, at least for large sample sizes. In situations where the large sample results does not hold or cannot be used for approximations due to lack of regularity, slow convergence or small sample sizes, one should be encouraged to look in other directions to estimate the mean squared error. This section suggests a few alternatives for estimating the mse using theory in the field of statistical resampling. Especially we will here propose schemes based on the famous bootstrap and its less computer intensive "father," the jackknife. For simplicity we will restrict ourselves to the situation of fully observed iid data.

5.5.1 Resampling techniques

Accuracy estimation based on resampling is a quite new technique in statistical analysis. It was a brilliant and almost magically well-working approach when it was invented. Nowadays the approach seems much more natural since most people got a computer which makes resampling a straightforward and quick task.

The most famous and widely used resampling technique of newer time is the bootstrap. It was invented by Efron in 1979 and as the name indicates, one really is lifting oneself up by one's own bootstraps by using only the data to learn about precisely the data. With the bootstrap one is able to estimate the uncertainty of a parameter estimator depending on data, without knowing its distribution. This is done by inserting an estimate of the distribution of data to get an approximate distribution also for the estimator. Nevertheless, deriving the distributional properties of an estimator is not always an easy task even if the distribution of the data is known. However, since most accuracy measures of interest can be represented as a function of an expectation, the bootstrap calls for the use of Monte Carlo integration[2] to approximate these attributes. The Monte Carlo version of the technique refers to what is known as bootstrapping, whereas calculating the true expectations under the assumed distribution of the data is often referred to as exact bootstrapping.

There are basically two types of bootstrapping, parametric and nonparametric bootstrapping. The former assumes a parametric distribution for the data and then samples directly from this distribution, and the latter assumes that the empirical distribution function is the true cdf of the data and thus samples from this distribution. Sampling from the distribution with cdf given by the ecdf is actually equivalent to sampling from the original data set with replacement. When one simply talks about bootstrapping the nonparametric version using the Monte Carlo method is usually what is meant, a convention we also will adopt.

The jackknife is an older estimation technique similar to the bootstrap. It was first invented by Quenouille in 1956 and further developed by Tukey two years later. The jackknife estimates the uncertainty of an estimator by measuring the sensitivity of leaving out observations and recalculating the estimator. The leave-one-out jackknife is the most common form of the jackknife. The variance and bias estimators based on this technique consist of comparisons between the estimators' values when each data point is left out consecutively to create new estimators based on the $n-1$ sized data sets. Such a comparison consists of estimator functions evaluated a total of n times. Hence, the leave-one-out jackknife is a much less computer intensive method than the bootstrap where a much greater number of function evaluations are necessary to obtain good estimators. The more general leave-d-out jackknife requires more computational power. The jackknife can actually be seen as an approximation to the bootstrap which is good for linear or close to linear estimators, but differs more for highly nonlinear estimators. For the latter case the jackknife has gained some criticism for its accuracy opposed to the bootstrap. This argument in addition to the fact that computing speed is not that big a concern with modern computers, has led to the convention that the bootstrap is often preferred by researchers. Nevertheless, the jackknife's advantage of giving the exact same result every time is still present and may be of importance in some applications. For more on the principles of the bootstrap and the jackknife, see e.g. Efron and Tibshirani (1993).

[2]Monte Carlo integration is a numerical integration techniques based on the law of large numbers.

5.5.2 Parametric bootstrapping in the limit

A nice property of the joint limiting distribution of relation (3.6) in lemma 3.1.2, is that it may be used as a basis for sampling pairs of $(\widehat{\mu}_{\rm np}, \widehat{\mu}_{\rm pm})^t$. One model selection approach to estimate the mse is then to use these samples to estimate the mse of the two estimators and hence constitute a model selection scheme. Before we go on to state such a routine, we stress that doing this is in some sense unnecessary since when the joint limiting distribution is used as an approximation for the finite n situation, we have already derived estimators of the mse with good properties. The routine is more or less stated for completeness and the fact that it matches the bootstrap paradigm. In addition the technique may be fruitful for situations less regular than those we consider here. Especially if one ought to consider loss functions which are less tractable analytically, the algorithm greatly simplifies estimation. We call this algorithm parametric bootstrap in the limit since we base the bootstrap on an approximation that is true in the limit experiment.

To justify the algorithm, we note that whenever relation (3.6) holds, also the following distribution holds approximately for large samples n:

$$\begin{pmatrix} \widehat{\mu}_{\rm np} \\ \widehat{\mu}_{\rm pm} \end{pmatrix} \doteq N_2\left(\begin{pmatrix} \mu_{\rm true} \\ \mu_{0,\rm pm} \end{pmatrix}, \frac{1}{n} \begin{pmatrix} V_{\rm np} & V_{\rm pm,np} \\ V_{\rm pm,np} & V_{\rm pm} \end{pmatrix} \right),$$

where $\doteq$ denotes approximately equally distributed. Now, using the usual estimates for the unknown quantities, we get the following computable approximation for the joint limiting distribution of the μ estimators:

$$N_2\left(\begin{pmatrix} \widehat{\mu}_{\rm np} \\ \widehat{\mu}_{\rm pm} \end{pmatrix}, \frac{1}{n} \begin{pmatrix} \widehat{V}_{\rm np} & \widehat{V}_{\rm pm,np} \\ \widehat{V}_{\rm pm,np} & \widehat{V}_{\rm pm} \end{pmatrix} \right). \tag{5.11}$$

Note that we in the above expression handle $\widehat{\mu}_{\rm pm}$ and $\widehat{\mu}_{\rm np}$ as given values and estimates of respectively $\mu_{0,\rm pm}$ and $\mu_{\rm true}$. Using this relation as a an approximation, yields the following parametric bootstrap routine:

1. Simulate a large number B of n-dimensional vectors $Y^{*,(b)} = (Y_1^{*,(b)}, \ldots, Y_n^{*,(b)})^t$, where $Y_i^{*,(b)}$ is a random sample from the joint normal distribution in (5.11).
2. Calculate the bootstrap estimates of the mses by using $\widehat{\mu}_{\rm np}$ as estimator for $\mu_{\rm true}$ as follows

$$\widehat{\rm mse}_{\rm lim.boot}(\widehat{\mu}_{\rm np}) = \frac{1}{B}\sum_{b=1}^{B} (\widehat{\mu}_{\rm np}^{*,(b)} - \widehat{\mu}_{\rm np})^2,$$

$$\widehat{\rm mse}_{\rm lim.boot}(\widehat{\mu}_{\rm pm}) = \frac{1}{B}\sum_{b=1}^{B} (\widehat{\mu}_{\rm pm}^{*,(b)} - \widehat{\mu}_{\rm np})^2.$$

Here $\widehat{\mu}_{\rm np}^{*,(b)}$ and $\widehat{\mu}_{\rm pm}^{*,(b)}$ are respectively the nonparametric and parametric μ estimators gotten by treating the data set $Y^{*,(b)}$ as the original data. As usual the FIC scheme consists of the mse estimators and selects the model with the estimator whose mse estimate is the smallest.

The above routine is stated under the assumption of only one parametric model. If there are several competing parametric models, one should simply run the simulations for each of the joint limiting distributions of $(\widehat{\mu}_{\rm np}, \widehat{\mu}_{\rm pm})^t$ and use the average of the mse estimates for the nonparametric estimators in place of $\widehat{\rm mse}_{\rm lim.boot}(\widehat{\mu}_{\rm np})$.

5.5.3 The bootstrap in the finite sample experiment

As mentioned in the introduction to this section, there are situations where basing mse estimators on large sample results cannot or should not be done. We now present a bootstrap routine that uses the bootstrap resampling technique to estimate the mse of each of the estimators without the need of any large sample results for the data set involved.[3] The proposed algorithm to arrive at mse estimators for the nonparametric and parametric μ estimators goes as follows:

1. Sample a large number B of n-dimensional vectors $Y^{*,(b)} = (Y_1^{*,(b)}, \ldots, Y_n^{*,(b)})^t$, where $Y_i^{*,(b)}$ is a random sample with replacement of the original data set $Y = (Y_1, \ldots, Y_n)^t$. Let $\widehat{\mu}_{\text{np}}^{*,(b)}$ and $\widehat{\mu}_{\text{pm}}^{*,(b)}$ for $b = 1, \ldots, B$ be respectively the nonparametric and parametric estimate of μ based on the b-th sampled set.

2. Calculate the bootstrap estimate of the mean squared error using $\widehat{\mu}_{\text{np}}$ as estimator for μ_{true}. For the nonparametric μ estimator, this corresponds to

$$\widehat{\text{mse}}_{\text{fin.boot}}(\widehat{\mu}_{\text{np}}) = \frac{1}{B}\sum_{b=1}^{B}(\widehat{\mu}_{\text{np}}^{*,(b)} - \widehat{\mu}_{\text{np}})^2,$$

   and for every parametric model with μ estimator $\widehat{\mu}_{\text{pm}}$, calculate

$$\widehat{\text{mse}}_{\text{fin.boot}}(\widehat{\mu}_{\text{pm}}) = \frac{1}{B}\sum_{b=1}^{B}(\widehat{\mu}_{\text{pm}}^{*,(b)} - \widehat{\mu}_{\text{np}})^2.$$

As usual the FIC scheme selects the model with smallest estimated mse. For small samples the bootstrapping procedure may also be carried out by exact bootstrapping. That involves calculating the μ estimates returned by each of the possible combinations of the sampled data. With a total number of $\binom{2n-1}{n}$ possible combinations, this is a manageable task for very small samples, but even at sample sizes as small as $n = 15$ there are 77 million evaluations that need to be performed. For $n = 30$ there are over $5 \cdot 10^{16}$ combinations which is obviously not reachable by any standard computer within reasonable time. The clever bootstrap technique of using random resamples of this set as an approximation is thus used in most situations. One must however be aware of the issue that when we resample at random the estimates of the mses will not be the same each time. The estimates will converge to an exact value when the number of samples $B \to \infty$, but for a finite B, two consecutive bootstrapping estimates of the same parameter will most certainly not be identical. As for bootstrapping in general, it is therefore very important that one uses a large number of samples B. Especially in the case where there are very little difference between the smallest estimated mses, one should be careful. If that is the case some additional resamples should be obtained to make sure that the winning model was not selected due to "luck" in the resampling procedure. In extreme cases of almost identical mse of different models, the uncertainty in the Monte Carlo procedure should be checked and possibly more efficient methods than just sampling at random should be performed. For more on such techniques, see any introductory book on the Monte Carlo Method, e.g. Rubenstein and Kroese (2008).

[3] To be precise, the bootstrap resampling technique uses Monte–Carlo simulations which are based on large sample results, but that concerns the number of resamplings, not the actual data set.

5.5.4 The jackknife in the finite sample experiment

A minor disadvantage of the bootstrap scheme of the previous subsection is, as emphasized, the inconsistency in the results. Even if an extra large B or wise resampling techniques would solve this problem in most situations, it is preferred that model selection criteria give consistent scores for each model. The jackknife solves this possible problem. The technique has the advantage over the bootstrap that it gives exactly the same results each time it is performed. Therefore this technique may be a more suitable approach to base model selection on.

There exists jackknife formulae for both the variance and the bias of quite general estimators. A problem with the bias formula is that it assumes some sort of closeness of the estimator and the estimand. Therefore, the usual formula does not work for the parametric estimator. As a result of this, we apply a different formula for estimation of the parametric uncertainty. The proposed algorithm to arrive at mse estimators for the nonparametric and parametric μ estimators goes as follows:

1. Let for $j = 1, \ldots, n$, $Y^*_{(j)} = (Y_1, \ldots, Y_{j-1}, Y_{j+1}, \ldots, Y_n)^t$ be the $(n-1)$-dimensional vectors consisting of all data points except the j-th. Let also $\widehat{\mu}^*_{\text{np},(j)}$ and $\widehat{\mu}^*_{\text{pm},(j)}$ be respectively the nonparametric and parametric estimates of μ based on the set $Y^*_{(j)}$. Let also $\overline{\widehat{\mu}^*_{\text{np}}} = (1/n)\sum_{j=1}^n \widehat{\mu}^*_{\text{np},(j)}$ and $\overline{\widehat{\mu}^*_{\text{pm}}} = (1/n)\sum_{j=1}^n \widehat{\mu}^*_{\text{pm},(j)}$

2. Calculate the jackknife estimates of the mse for $\widehat{\mu}_{\text{np}}$ and $\widehat{\mu}_{\text{pm}}$ as follows:

$$\widehat{\text{mse}}_{\text{jack}}(\widehat{\mu}_{\text{np}}) = (n-1)^2\left(\overline{\widehat{\mu}^*_{\text{np}}} - \widehat{\mu}_{\text{np}}\right)^2 + \frac{n-1}{n}\sum_{j=1}^n\left(\widehat{\mu}^*_{\text{np},(j)} - \overline{\widehat{\mu}^*_{\text{np}}}\right)^2,$$

$$\widehat{\text{mse}}_{\text{jack}}(\widehat{\mu}_{\text{pm}}) = \left(\overline{\widehat{\mu}^*_{\text{pm}}} - \widehat{\mu}_{\text{np}}\right)^2 + \frac{n-1}{n}\sum_{j=1}^n\left(\widehat{\mu}^*_{\text{pm},(j)} - \overline{\widehat{\mu}^*_{\text{pm}}}\right)^2.$$

The term $(n-1)$, included in all formulae except the parametric bias formula, adjusts for the fact that removing just one single data point will give values very close to $\widehat{\mu}_{\text{np}}$. The squared bias estimator for the parametric estimator is created by intuition and testing, and seems to work out fairly well. Note also that we have not adjusted for possibly overestimation caused by squaring the bias estimator. Such may be performed by e.g. another bootstrap to estimate the variance of the bias estimator. This is however omitted here since the main reason for using the jackknife was to get a single value not dependent on the number of simulations we run. As usual a FIC apparatus is carried out by selecting the model whose above mse estimate is the smallest.

5.5.5 Additional notes

We have emphasized earlier that the schemes based on resampling in the finite n experiment is a valid alternative when e.g. samples are too small to rely on large sample approximations of the main scheme of this chapter. Care must however be taken regarding the resampling schemes even if they do not use large sample approximations. One must be aware that resampling the data in any way can only extract as much information from the data as it possess. Thus, when little information is available through the data, not only will the uncertainty in the actual μ estimators be quite large, but also the uncertainty in the resulting mse estimates will be

quite large. So even if resampling is considered a good technique for small samples, there is considerable uncertainty involved also when such techniques are applied.

The smoothed bootstrap is a variant of the bootstrap that is often used for small samples. This technique first smooth each data point using a kernel smoother. Then one samples from this smoothed distribution instead of resampling the data with replacement. Especially for focus parameters like the median or any other quantile one might argue that such a technique will reduce the uncertainty greatly simply because one will no longer get a great number of identical estimates of μ in the resampling step.

The schemes presented in this section do not only differ in terms of estimation technique, but also on exactly what is being estimated. We have previously been splitting the mean squared error into terms of squared bias and variance, and estimated these separately. On the other hand, this section used only estimates of the mse directly. The simple reason for directly estimation in this section, is that we can. We could however have divided the mse into the terms of squared bias and variance also here. Such splitting may yield slightly different results.

Note finally that using resampling techniques on only the squared bias term using the standard plug-in estimators based on large sample results for the variance, or vice versa, also yields a perfectly valid FIC scheme.

5.6 Parametric models with other convergence rates

As mentioned in section 3.3 the FIC scheme of chapter 3 does not support model selection where the uniform distribution $U[0,\theta]$ is one of the parametric models. The reason it does not fit the scheme is that the ML estimator $\widehat{\theta}_n = \max_{i=1,\ldots,n}\{Y_i\}$ converges towards θ_0 with rate n, rather than the usual $\sqrt{n}$. Despite this unpleasant behavior one should be able to perform model selection between nonparametrics and parametrics also for this situation.

As shown in Lehmann (1998, Example 2.3.7), $n(\widehat{\theta}_n - \theta_0) \xrightarrow{L} \text{Exp}(1/\theta_0)$. Thus, by the delta method (theorem B.2.8) (which works fine even for convergence rates other than $\sqrt{n}$) we get that

$$n(\widehat{\mu}_{\text{pm}} - \mu_{0,\text{pm}}) \xrightarrow{L} \left.\frac{\partial \mu_F}{\partial \theta}\right|_{\theta_0} \text{Exp}\left(\frac{1}{\theta_0}\right).$$

Thus, for large n the variance of $n\widehat{\mu}_{\text{pm}}$ may be approximated by

$$V_{\text{pm}} = \left(\left.\frac{\partial \mu_F}{\partial \theta}\right|_{\theta_0}\right)^2 \theta_0^2,$$

since the variance of an exponential distributed random variable $\text{Exp}(\lambda)$ is $1/\lambda^2$. V_{pm} may consequently be estimated by inserting $\widehat{\theta}_n$ for θ_0, i.e.

$$\widehat{V}_{\text{pm}} = \left(\frac{\partial \mu_F}{\partial \theta}|_{\widehat{\theta}_n}\right)^2 \widehat{\theta}_n^2. \tag{5.12}$$

As a result, the variance of $\widehat{\mu}_{\text{pm}}$ may be estimated by $\frac{1}{n^2}\widehat{V}_{\text{pm}}$. Since $\widehat{\mu}_{\text{np}}$ and its mse estimates are not affected by the parametric distributions, the variance of $\widehat{\mu}_{\text{np}}$ may as usual be estimated by $\frac{1}{n}\widehat{V}_{\text{np}} = \frac{1}{n}\widehat{\nu}$, and the squared bias simply by zero. The bias of $\widehat{\mu}_{\text{pm}}$ may be estimated by $\widehat{b} = \widehat{\mu}_{\text{pm}} - \widehat{\mu}_{\text{np}}$. The usual strategy of estimating the squared bias by squaring the estimate and subtracting an estimator for its variance, is however a bit troublesome. This is problematic

because it is hard to estimate the covariance between $\widehat{\mu}_{\text{pm}}$ and $\widehat{\mu}_{\text{np}}$. The reason for this is that $\widehat{\mu}_{\text{pm}}$ or even $\widehat{\theta}_n = \max_{i=1,\ldots,n}\{Y_i\}$ cannot be written as a sum of iid variables in a convenient way, and therefore the usual sample covariance estimator cannot be applied directly. There are however alternative estimation techniques available. One of them is bootstrapping as introduced in section 5.5. Assume now that we were able to estimate the variance of $\widehat{b}$ by $\widehat{V}_{\text{b}*}$ using some estimation technique. We may then propose the following FIC scheme:

$$\begin{aligned}
\text{FIC}(\widehat{\mu}_{\text{np}}) &= \frac{1}{n}\widehat{V}_{\text{np}}, \\
\text{FIC}(\widehat{\mu}_{\text{pm}}) &= (\widehat{\mu}_{\text{pm}} - \widehat{\mu}_{\text{np}})^2 - \widehat{V}_{\text{b}*} + \frac{1}{n^2}\widehat{V}_{\text{pm}},
\end{aligned}$$

where $\widehat{V}_{\text{pm}}$ is as given in equation (5.12) and $\widehat{V}_{\text{np}} = \widehat{\nu}$. As usual, a FIC scheme of this type will select the model with the smallest FIC score. As an illustration, consider the simple case where the focus parameter is the expectation $\mu = E_G[Y_i]$. We then get $\widehat{\mu}_{\text{np}} = \overline{Y} = 1/n\sum_{i=1}^n Y_i$ and $\widehat{\mu}_{\text{pm}} = \widehat{\theta}_n/2$, which by some algebra give the following FIC scores:

$$\begin{aligned}
\text{FIC}(\widehat{\mu}_{\text{np}}) &= \frac{1}{n}\widehat{\sigma}^2, \\
\text{FIC}(\widehat{\mu}_{\text{pm}}) &= \left(\frac{\widehat{\theta}_n}{2} - \overline{Y}\right)^2 - \widehat{V}_{\text{b}*} + \frac{1}{n^2}\widehat{\theta}_n^2,
\end{aligned}$$

where $\widehat{\sigma}^2$ is the sample variance. The same strategy may of course be applied to other parametric distributions with ML estimators with other convergence rates than $\sqrt{n}$, provided that we are able to derive the precise limiting distribution of its ML estimator.

Chapter 6

Weighted FIC

Pure focused model selection considers a single, fully specified focus parameter μ and is what we have directed our attention to so far. However, as mentioned when discussing why the focus parameter cannot be multidimensional in section 3.3, one may consider a model selection criterion optimized for estimating a set of focus parameters. This chapter presents a general model selection strategy for such situations. We start out by deriving a quite general scheme of this type. Furhtermore, the relation to a certain goodness of fit test is discussed and some properties of this test are shown. In the end we illustrate the regular iid version of the scheme on a real data example.

We will denote the weighted FIC scheme of this chapter by wFIC.[1] In the spirit of this thesis, we will choose among the usual nonparametric model and a number of parametric models. As mentioned, the derivation in this chapter is kept fairly general and not restricted to a certain type of data. Basically we assume no more than that there exist a FIC scheme that applies to all focus parameters for which the weight function assigns positive weight.

Why wFIC schemes may be useful can be seen from different points of view. When the interest is wider than just a single fully specified focus parameter, the radical FICologist would suggest performing individual model selection for each of the focus parameters. For a rather small discrete set this is highly applicable, but for larger discrete and also continuous sets of focus parameters, one either has to apply some kind of approximation, or do a wider theoretical analysis exploring when the criterion changes winning model. Following such a strategy may thus be a quite comprehensive task in some situations. In many situations it is beneficial to base different types of inference on the same model. Thus, the strategy of wFIC may be fruitful when the set of focus parameters is not the simplest, and for sure in situations where the interest of the model fitting is not completely specified prior to the analysis. Analysis where the task is to estimate say the vague "upper quantile" or the even more unclear "middle of the distribution", may be seen as typical examples of the latter. In other situations, some focus parameters may be of greater importance than others. Attaching a weight function to the set of focus parameters specifying the importance of each of them relative to the others and applying wFIC, is a possible solution to all of these problems. The overall model selectors AIC and BIC chooses the model that is generally best for estimation based on the model, whereas wFIC remains focused in the way that zero weight are given to uninteresting estimation purposes, while exact specification of a single focus is not required. In that point of view, wFIC may be

[1] The idea has its origin from Claeskens and Hjort (2008) where wFIC or AFIC (for Average FIC) as it is termed by these authors, does a similar job for model selection among parametric models.

seen as a golden mean.

6.1 A general derivation of wFIC

So far in this thesis, the proposed model selection criteria have had one goal in common: To choose the model minimizing the mean squared error of the model based estimators of a focus parameter μ. For all of these the strategy has consisted of estimates of this mean squared errors and the model with the smallest estimate has been chosen. This strategy may be seen as a search for the model minimizing a certain risk function. The risk function is in this case defined as the expectation of the loss function

$$L(\widehat{\mu}) = (\widehat{\mu} - \mu_{\text{true}})^2 .$$

Schemes aiming on risk minimization for other loss functions may of course also be appropriate. As mentioned in the above introduction, we are here interesting in a scheme performing well for a set of focus parameters, where their importance is specified in terms of a weight function. We denote by $W(u)$ the weight function associated with the focus parameter $\mu(u)$. The weight assigned to $\mu(u)$ is then determined by the function $w(u)$ which have $W(u)$ as cdf. For the set of focus parameters of interest given by $\mu(u)$ for some $u \in \mathbb{R}$, it is natural to define the following loss function:

$$L_W(\widehat{\mu}, W) = \int (\widehat{\mu}(u) - \mu_{\text{true}}(u))^2 \, \mathrm{d}W(u).$$

The risk of this loss function is consequently given by

$$\text{risk}_W(\widehat{\mu}, W) = E[L_W(\widehat{\mu}, W)].$$

One could possibly think of weight functions that are random, e.g. depending on data, but a fixed weight function without randomness is flexible enough to handle most practical situations. For such situations, the above risk function simplifies to

$$\text{risk}_W(\widehat{\mu}, W) = \int E_G \left[(\widehat{\mu}(u) - \mu_{\text{true}}(u))^2 \right] \, \mathrm{d}W(u) = \int \text{mse}(\widehat{\mu}(u)) \, \mathrm{d}W(u). \tag{6.1}$$

The model selection scheme of wFIC will thus aim at estimating equation (6.1) for estimators based on the different candidate models. Estimating this expression may at first sight look rather difficult, but for situations where a FIC scheme is already developed, it is quite straight forward. Assume now that a there exist FIC schemes that estimates the mse for all values of u with positive weight, and all schemes include the same candidate models. Then, we get the following risk estimates

$$\text{wFIC}(\widehat{\mu}_M) = \widehat{\text{risk}}_W(\widehat{\mu}, W) = \int \widehat{\text{mse}}(\widehat{\mu}_M(u)) \, \mathrm{d}W(u) = \int \text{FIC}(\widehat{\mu}_M(u)) \, \mathrm{d}W(u), \tag{6.2}$$

for each model M in the set of candidate models. For a set of focus parameters indexed by $\mu(u)$ and a corresponding weight function $W(u)$ analytical expressions can thus be obtained by integration. Especially, for the case of a discrete weight function, the integrals reduces to sums, and the estimated risk from equation (6.2) reduces further to

$$\text{wFIC}(\widehat{\mu}_M) = \sum_{u \in \mathcal{U}} \text{FIC}(\widehat{\mu}_M(u)) w(u), \tag{6.3}$$

for $\mathcal{U}$ the set of u values with nonzero weight. For a continuous set $\mathcal{U}$, where direct integration is either hard or impossible, numerical integration may be performed to calculate the wFIC scores. It is however even simpler to discretize the weights to a finite set of u values and use equation (6.3) as an approximation in those cases. By increasing the number of u values, this method can be made arbitrarily precise. Note also that the situation where there is only one single u value with nonzero weight reduces to the usual FIC scheme, as desired.

In some of the situations treated in this chapter, we have presented an alternative, or modified FIC formula where any possibly negative estimates of a positive quantities are set to zero. Also for this general wFIC scheme, such a modification may be appropriate. Negative estimates of squared quantities in the FIC scheme regard the parametric estimators only, and will thus regard only parametrics here as well. For a general parametric μ estimator of a FIC scheme, as usual denoted by $\widehat{\mu}_{\mathrm{pm}}$, we write

$$\mathrm{FIC}(\widehat{\mu}_{\mathrm{pm}}) = \widehat{b}^2(\widehat{\mu}_{\mathrm{pm}}) - \widehat{V}_{\mathrm{b}} + \widehat{V}_{\mathrm{pm}},$$

and consequently the wFIC score of the same $\widehat{\mu}_{\mathrm{pm}}$ may be written

$$\mathrm{wFIC}(\widehat{\mu}_{\mathrm{pm}}) = \int \widehat{b}^2(\widehat{\mu}_{\mathrm{pm}}(u))\,\mathrm{d}W(u) - \int \widehat{V}_{\mathrm{b}}(u)\,\mathrm{d}W(u) + \int \widehat{V}_{\mathrm{pm}}(u)\,\mathrm{d}W(u).$$

Hence, motivated by modifications to the FIC scheme, we present the following modified wFIC scheme

$$\begin{aligned}
\mathrm{wFIC}^*(\widehat{\mu}_{\mathrm{np}}) &= \mathrm{wFIC}(\widehat{\mu}_{\mathrm{np}}) = \int \widehat{V}_{\mathrm{np}}(u)\,\mathrm{d}W(u), \\
\mathrm{wFIC}^*(\widehat{\mu}_{\mathrm{pm}}) &= \left\{ \int \widehat{b}^2(\widehat{\mu}_{\mathrm{pm}}(u))\,\mathrm{d}W(u) - \left[\int \widehat{V}_{\mathrm{b}}(u)\,\mathrm{d}W(u) \right]^+ \right\}^+ + \int \widehat{V}_{\mathrm{pm}}(u)\,\mathrm{d}W(u).
\end{aligned}$$

Note that the positive parts are taken after integration. Using this scheme is therefore not equivalent to simply weighting the adjusted FIC scores according to the weight function W. The motivation for using this type of modification as opposed to inserting the adjusted FIC scores directly is that we are now not aiming on estimating the mse for each $u \in \mathcal{U}$, but rather on estimating the expected loss as given in equation (6.1). Thus, the estimator should strive at estimating this quantity as exactly as possible. This is most naturally done by making sure that the integrals are positive. Note however that there is no strict rule forbidding one to adjust the mse estimators before integrating. Doing that will lead to a valid estimator of the same quantity. The point is rather that adjusting post integration is, as argued above, the natural way of doing this. See also a similar discussion in Claeskens and Hjort (2008, chapter 6.9).

It should be noted that not all natural weight functions corresponds to a cdf. In the situation where the focus parameter is the cdf, one may want to weight all the evaluation points of an unbounded cdf equally. This corresponds to the weight function $W(u) = cu$ which is not a cdf for an unbounded set of u values. Such situations may however be approximated sufficiently well by restricting the evaluation set for the weight function to a bounded interval $[a, b]$, since W then is a cdf. This approximation can be made arbitrarily good by increasing the interval, and will therefore cover any practical situation.

6.2 wFIC as a goodness of fit test

As mentioned in section 2.2 there exists quite a few goodness of fit tests. The quantity

$$n \int \left(\widehat{G}_n(u) - H(u) \right)^2 \, \mathrm{d}W(u), \tag{6.4}$$

for some nondecreasing weight function W and some fixed cdf H, where proposed independently in Cramér (1928) and von Mises (1931) as a measure of the goodness of fit for the distribution with cdf H. The modification of this function where $\mathrm{d}W(u) = w(H(x)) \, \mathrm{d}H(x)$, gives a distribution free measure that is common to use for goodness of fit testing. This measure is often referred to as the Cramér–von Mises goodness of fit measure. As we will see, there is a clear connection between goodness of fit testing of $H = F_{\theta_0}$ based on equation (6.4) and a certain wFIC scheme.

Consider now the situation with iid data $Y_1, \ldots, Y_n$ stemming from some fixed, but unknown distribution with cdf G. Furthermore, assume we are interested in estimating the cdf well at all points y_0, where well is determined by the weight function which is given by the cdf W. The two superior estimation techniques for this task are estimation directly from the empirical distribution function, or going via ML estimation for some parametric distribution function and base estimation on the cdf of this fitted parametric distribution. For simplicity let us assume that we fit only one parametric distribution. The proposed wFIC scheme is a natural approach to choose which of the two estimation techniques to base further inference on. For this situation the wFIC scores may be written as

$$\mathrm{wFIC}(\widehat{\mu}_M) = \int \mathrm{FIC}(\widehat{\mu}_M(u)) \, \mathrm{d}W(u),$$

for $M = \mathrm{np}$ and $M = \mathrm{pm}$. Inserting the FIC formulae for $\mathrm{FIC}(\widehat{\mu}_{\mathrm{np}})$ and $\mathrm{FIC}(\widehat{\mu}_{\mathrm{pm}})$, we get that the nonparametric approach is chosen whenever

$$Z_n^2 = \int n \left(\widehat{G}_n(u) - F(u; \widehat{\theta}_n) \right)^2 \mathrm{d}W(u) \geq 2 \int \left(\widehat{V}_{\mathrm{np}}(u) - \widehat{V}_{\mathrm{pm,np}}(u) \right) \mathrm{d}W(u). \tag{6.5}$$

The left hand side of this inequality is just the same as the goodness of fit measure of equation (6.4). Thus, this type of wFIC scheme may be seen as testing the goodness of fit of the distribution with cdf $F(y; \widehat{\theta}_n)$.

Consider now the hypothesis test of $H_0 : G = F_{\theta_0}$ against the two-sided alternative $H_A : G \neq F_{\theta_0}$, which rejects when inequality (6.5) is fulfilled. I.e. we insert $F_{\widehat{\theta}_n}$ for the unknown F_{θ_0} and applies the wFIC scheme. The level of this test is the probability that inequality (6.5) is fulfilled under the null hypothesis. Especially we will consider the asymptotic level via large sample theory.

Note now that when $\widehat{V}_{\mathrm{np}}(u)$ and $\widehat{V}_{\mathrm{pm,np}}(u)$ are consistent not only for every u, but uniformly on the set where the weight function assigns positive weight, we get that the right hand side of inequality (6.5) converges almost surely to $2 \int (V_{\mathrm{np}}(u) - V_{\mathrm{pm,np}}(u)) \, \mathrm{d}W(u)$. This follows from the continuous mapping theorem (B.2.9), since integration is a continuous operation and W is a cdf and consequently also a measure. When this holds equation (6.5) is asymptotically equivalent to

$$Z_n^2 = \int n \left(\widehat{G}_n(u) - F_{\widehat{\theta}_n}(u) \right)^2 \mathrm{d}W(u) \geq 2 \int (V_{\mathrm{np}}(u) - V_{\mathrm{pm,np}}(u)) \, \mathrm{d}W(u). \tag{6.6}$$

Furthermore, an intermediate result in Durbin (1973) may be restated as $X_n(u) = \sqrt{n}(\widehat{G}_n(u) - F(u;\widehat{\theta}_n))$ converges weakly to a certain zero-mean Gaussian process.[2] Thus, it follows from the continuous mapping theorem that since W is a cdf and then also a measure, Z_n^2 converges in law to a certain random variable Z^2 which may be represented as an integral over a squared zero-mean Gaussian process. Careful mathematical treatment of Z^2 would lead to a fully computable expression for the limiting distribution of Z_n^2. Since such process theory is beyond the scope of this thesis, we will not handle this. Instead, we use simulations to investigate the limit behavior of inequality (6.6) and consequently arrive at the asymptotic level of such a test.

6.2.1 Simulations

To use simulation to investigate the behavior of Z_n^2 is an alternative approach to the more technical and direct analytical derivation. Simulating with a weight function that assigns equal weight to a finite number of values on a bounded interval, makes comparison with the most common goodness of fit tests the smoothest. Consider the following routine: *For a large number n (1000 seems to be sufficient), simulate a number of iid data sets with length n from some parametric distribution, and then calculate $X_n(u)^2 = n\left(\widehat{G}_n(u) - F(u;\widehat{\theta}_n)\right)^2$ for many different u values on some interval. For each simulation, the mean of the $X_n(u)^2$ values (denoted $\mathcal{Z}_n^2$) is an approximate sample from the limiting distribution of Z_n^2. For each simulation, check if inequality (6.6) is fulfilled or not when $\mathcal{Z}_n^2$ replaces Z_n^2. The proportion of the times the inequality is fulfilled then estimates the asymptotic level of the test for exactly this parametric distribution.*

The first situation we will investigate is probably the most natural situation on could think of, namely normality. For this particular situation, we take the standard normal distribution $N(0,1)$ as true and sample 10^6 copies of iid data sets from this distribution with $n = 1000$. We also take u values on the interval $[-3.1, 3.1]$ with steps of size 0.01. A standard normal distributed variable has probability less than 0.002 to be outside this interval. This simulation study show rather surprisingly that the asymptotic level is very close to the commonly, but rather artificially chosen significance level 0.05. With 4 valid digits, the simulation gives an approximate asymptotic level of 0.0495.

Similar simulations indicated that the convergence of Z_n^2 is rather fast, as sample sizes as small as $n = 5$ give almost the same results. Note however that for such small sample sizes the right hand side of inequality (6.5) would vary much more, so the finite n level of this test would probably differ a bit from this. The refinement and span of the u values could clearly also matter here. However, tests show that smaller steps between the u values or greater span of the interval gives the same results.

Secondly, we consider the exponential distribution. We take the standard exponential distribution with parameter value $\theta = 1$ as true and sample also here 10^6 copies of iid data sets from this distribution with $n = 1000$. We take also assign equal weight to all values on the interval $[0, 6.3]$ with steps of size 0.01. A standard exponential distributed variable has probability less than 0.002 to be outside this interval. For this parametric family the simulations show that the asymptotic level is somewhat larger than for the normal distribution. With 4 valid digits, the simulations result in an approximate asymptotic level of 0.0698.

The above simulation studies give the answer for wFIC which is analogous for the selection probability of 0.157 for FIC as derived in section 3.5. Furthermore, these results show the

[2] Weak convergence may be seen as the general metric space analogue of convergence in law.

connection between a certain special case of the wFIC scheme, for the regular iid data situation, and goodness of fit tests. The asymptotic level for the wFIC tests are smaller than those of FIC, indicating that parametrics will be chosen more often. Note however that for other parametric families it is likely that the asymptotic levels are different from the ones obtained here.

6.2.2 Parameter independence

The obtained asymptotic level of the two above situations is based only on one particular choice of parameter values. By letting the weight function vary naturally with the parameters values or estimates of them, one may show that the asymptotic level is independent of the actual parameter values for a wide class of such tests. For simplicity we will here concentrate on the situation with a continuous weight function W. The class of parametric models we shall show parameter independence for, is the so-called location scale family. The distributions of this family has the property that if X is a random variable from the family with expectation ξ and variance σ^2, then $(X-\xi)/\sigma$ is the zero-mean unit-variance member of the family. We shall use the following property of such families:

$$F(y;\xi,\sigma) = F\left(\frac{y-\xi}{\sigma};0,1\right).$$

Note now that the ecdf may be written on a similar form. By letting $\widehat{G}_{n,\text{unit}}(y)$ be the cdf of the data $X_i = \frac{Y_i-\widehat{\xi}}{\widehat{\sigma}}$, for $i=1,\ldots,n$, we get that

$$\widehat{G}_n(y) = \widehat{G}_{n,\text{unit}}\left(\frac{y-\widehat{\xi}}{\widehat{\sigma}}\right),$$

where indeed $\widehat{G}_{n,\text{unit}}$ is independent of the estimated parameters. Furthermore, let the weight function depend on the parameters of the parametric distribution such that also

$$W(y;\xi,\sigma) = W\left(\frac{y-\xi}{\sigma};0,1\right).$$

Since W is continuous, it follows that

$$\mathrm{d}W(y;\xi,\sigma) = \frac{1}{\sigma}w\left(\frac{y-\xi}{\sigma};0,1\right)\,\mathrm{d}y,$$

where $w(y;\xi,\sigma)$ is the density of $W(y;\xi,\sigma)$. Note also that if a and b are the bounds of W (they need not to be finite), we have

$$0 = W(a;0,1) = W(a\sigma+\xi;\xi,\sigma), \qquad 1 = W(b;0,1) = W(b\sigma+\xi;\xi,\sigma).$$

Now, by letting the weight function depend on the estimated values of ξ and σ, which here are the ML estimates $\widehat{\theta}_n = (\widehat{\xi},\widehat{\sigma})^t$, we get that

$$\begin{aligned}
Z_n^2 &= n\int_{a\widehat{\sigma}+\widehat{\xi}}^{b\widehat{\sigma}+\widehat{\xi}} \left(\widehat{G}_n(u) - F(u;\widehat{\xi},\widehat{\sigma})\right)^2 \mathrm{d}W(u;\widehat{\xi},\widehat{\sigma}) \\
&= n\int_{a\widehat{\sigma}+\widehat{\xi}}^{b\widehat{\sigma}+\widehat{\xi}} \left(\widehat{G}_{n,\text{unit}}\left(\frac{u-\widehat{\xi}}{\widehat{\sigma}}\right) - F\left(\frac{u-\widehat{\xi}}{\widehat{\sigma}};0,1\right)\right)^2 \frac{1}{\widehat{\sigma}}w\left(\frac{u-\widehat{\xi}}{\widehat{\sigma}};0,1\right)\,\mathrm{d}u \\
&= n\int_a^b \left(\widehat{G}_{n,\text{unit}}(s) - F(s;0,1)\right)^2 w(v;0,1)\,\mathrm{d}s.
\end{aligned}$$

In the last equality we performed a change of integration variable to $s = (u - \widehat{\xi})/\widehat{\sigma}$. The last expression is independent of the parameters. As a consequence, so is Z_n^2 for each n. Note that even if we did not open for weight functions depending on data earlier in the chapter, the choice of weight function is not problematic here, since, as we see from the above equation, the integral may be rewritten to be independent of the parameters that depends on the data.

Using similar arguments one may also show that the right hand side of inequality (6.5) is parameter independent as well, possibly under some additional assumptions making sure that the convergence of $\widehat{V}_{\text{pm,np}}(u)$ is uniform. Since we do not go into these arguments in this thesis, we restrict ourselves to indicate parameter independence in the assumed limit instead, i.e. for the right hand side of inequality (6.6). Note that $V_{\text{np}}(u) = G(u)(1 - G(u))$, and that $V_{\text{pm,np}} = V_{\text{pm}}$ since we are working under the null hypothesis. As a result, any dependence between $(\xi, \sigma)^t$ and $V_{\text{np}}(u) - V_{\text{pm}}(u)$ are through $G(u)$ and $F(u; \xi, \sigma)$. One may therefore write $V_{\text{np,unit}}(u)$ and $V_{\text{pm,unit}}(u)$ for the unit versions of $V_{\text{np}}(u)$ and $V_{\text{pm}}(u)$. We therefore get

$$
\begin{aligned}
&2 \int_{a\sigma+\xi}^{b\sigma+\xi} \left(V_{\text{np}}(u) - V_{\text{pm}}(u)\right) \, \mathrm{d}W(u; \xi, \sigma) \\
&= 2 \int_{a\sigma+\xi}^{b\sigma+\xi} \left(V_{\text{np,unit}}\left(\frac{u-\xi}{\sigma}\right) - V_{\text{pm,unit}}\left(\frac{u-\xi}{\sigma}\right)\right) \frac{1}{\sigma} w\left(\frac{u-\xi}{\sigma}; 0, 1\right) \, \mathrm{d}u \\
&= 2 \int_a^b \left(V_{\text{np,unit}}\left(s\right) - V_{\text{pm,unit}}\left(s\right)\right) w\left(v; 0, 1\right) \, \mathrm{d}s,
\end{aligned}
$$

which ensures parameter independence also for the right hand side of inequality (6.6). As a consequence, the test is parameter independent for any location scale family provided the weight function is chosen as indicated. It is well-known that the normal distribution belongs to the location scale family, so the result holds for testing normality. The exponential distribution is not a member of this family, but the same result follows also for this distribution by letting only $\sigma = 1/\theta$ vary and simply ignoring the location adjustment.

Remark 4. *The connection between wFIC and goodness of fit testing for the regular iid situation, may hold also for censored data. It is immediate that inequalities corresponding to inequality* (6.5) *may be stated for the FIC schemes in chapter 4. In such cases the cumulative hazard rate or the survival function replaces the cdf. In particular Hjort (1990) studies the larges sample behavior of* $\sqrt{n}\left(\widehat{A}_{\text{np}}(t) - \widehat{A}_{\text{pm}}(t)\right)$ *as a process. By using these results, and carefully carrying out the integration of such a process, one should arrive at similar results as above also in the censored setting.*

6.3 An example of wFIC in use

In the first section of this chapter we treated wFIC in a totally general matter, making it possible to use for a great number of situations. It is rather straight forward to carry out the formulae in each of the situations we have derived matching FIC schemes for earlier in the thesis. To get a feeling on how the scheme works, we here give an example of the regular iid version in use. We will focus on the data set considered in section 3.9.1 of chapter 3, consisting of a measure of the number of steps taken on a daily basis by 3464 Norwegians.

In the mentioned example, we were interested in the proportion of the population that fulfilled the recommendation of at least 10 000 daily steps from the Norwegian Directorate

of Health. Here we call attention against the most enthusiastic walkers. Suppose our task is to find a model that models the number of steps taken by the people that walk the most. Specifically, we will define the set of focus parameters with nonzero weight as the discrete set of quantiles from 0.91 to 0.99, with steps of size 0.01. Furthermore, we weight the set of focus parameters according to the following weight function gotten when:

$$w(u) = (5 - |0.95 - u|)/25, \quad u = 0.91, 0.92, \ldots, 0.99,$$

and $W(u) = \sum_{s \leq u} w(s)$. This means that most weight will be given to the 0.95-quantile and that the weight will be decreasing linearly from this value to both the 0.99 and the 0.91 quantile. All other quantiles are given zero weight. Applying the modified FIC scheme above gives results as given in table 6.1.

	dim	$\widehat{\text{bias}^*}$	$\widehat{\text{sd}}$	$\widehat{\text{wRMSE}}$	Rank
Nonpar	Inf	0	174.77	174.77	2
Normal	2	531.89	123.38	546.01	3
Skewed normal	3	0	145.46	145.46	1
Lognormal	2	873.36	188.80	893.53	4

Table 6.1: Results from the wFIC scheme when focus is on the greatest quantiles $0.91, \ldots, 0.99$.

In corresponding example in chapter 3, the log-normal distribution was selected as the best model for estimating the proportion of the population that walk more than 10 000 steps per day. As seen from the results in table 6.1, the lognormal model does not perform well at all when estimating the largest quantiles of the distribution. The skewed normal distribution wins in terms of this modified wFIC scheme. The second best nonparametric model performs quite well also, but the two others are as seen not good models at all. That the skewed normal distribution is the winner here, is maybe not a big surprise if one considers the density plot given in figure 3.3. The skewed normal distribution seems to fit very well to data especially for the largest values, which is what we focus on here. Table 6.2 gives the estimates of the quantiles from 0.91 to 0.99 based on the winning skewed normal model. As seen, the model estimates that the 5 percent most active walkers in the adult Norwegian population walk a total of 13442 steps per day. Furthermore, the model predicts that the upper 1 percent of the population walk more than 6000 steps more than recommended.

u	0.91	0.92	0.93	0.94	0.95	0.96	0.97	0.98	0.99
$\widehat{\mu}_{\text{winner}}(u)$	12338	12569	12824	13111	13442	13833	14318	14970	16011

Table 6.2: Resulting estimates of the greatest quantiles based on the winning skewed normal distribution.

Chapter 7

Model Averaging

Model selection uses data to select one model among a set of candidate models which according to a chosen criterion, is the best. Even if the scheme indicated that the chosen model was only marginally better than others, only the winning model is used for further inference. The theme of this chapter is that of model averaging which base further inference not only on one single model, but may use several models to e.g. estimate a certain quantity.

We will start this chapter by giving a rough introduction to model averaging. Then we will state a model averaging scheme in the spirit of this thesis. Moreover, we will derive the limiting distribution for this model averaging estimator under a few assumptions and discuss its usefulness. Finally, we will apply the estimator to a real data example.

7.1 Introduction to model averaging

Model averaging take advantage of the fact that several models may be suitable for a given data set. Let $\mathcal{M}$ be the set of candidate models for a certain data set. The model averaging estimator of a parameter μ is then given by

$$\widehat{\mu}_{\text{final}} = \sum_{M \in \mathcal{M}} W(M)\widehat{\mu}_M, \tag{7.1}$$

for some weight function W,[1] where $\sum_{M\in\mathcal{M}} W(M) = 1$. For example one may consider the situation where the estimates from each model are given equal weight:

$$\widehat{\mu}_{\text{avg}} = \sum_{M \in \mathcal{M}} \frac{\widehat{\mu}_M}{|\mathcal{M}|},$$

where $|\mathcal{M}| = \#\{M \in \mathcal{M}\}$ is the number of candidate models. Even if one may consider such estimators where the weights are nonrandom, the most interesting cases includes a random weight W. Note also that by using indicator weights, one may write

$$\widehat{\mu}_{\text{IC}-\text{winner}} = \sum_{M \in \mathcal{M}} \mathbf{1}_{\{M=M_{\text{IC}-\text{winner}}\}}(M)\widehat{\mu}_M, \tag{7.2}$$

[1] Not to be confused with the weight function in the previous chapter.

where $M_{\text{IC}-\text{winner}}$ denotes the winner of a certain model selection scheme and $\widehat{\mu}_{\text{IC}-\text{winner}}$ denotes the μ estimator under this model. Thus, model averaging may be seen as a generalization of model selection.

Model averaging is mostly used with weights determined by some model selection scheme. As mentioned in the introduction to BIC in section 2.2, the BIC scores are actually the decisive ingredient in an approximation to the posterior model selection probability when using a flat prior. Reversing the formula, one gets an approximating formula for the posterior model selection probability given by

$$\widehat{Pr}\{M|\text{ data}\} = \frac{\exp\left(\frac{1}{2}\text{BIC}(M)\right)}{\sum_{M'\in\mathcal{M}}\exp\left(\frac{1}{2}\text{BIC}(M')\right)}. \tag{7.3}$$

I.e. when considering all these candidates equally likely (and no other models are possible) before data is considered, formula (7.3) estimates the probability that model M is true. Thus, using these probabilities as weights for each of the models, $\widehat{\mu}_{\text{final}}$ is the expected posterior value of μ under this prior assumption. This scheme is often referred to as the smooth BIC scheme. Buckland et al. (1997) suggest using an analogous weight function based on AIC, given by

$$W_{\text{AIC}'}(M) = \frac{\exp\left(\frac{1}{2}\text{AIC}(M)\right)}{\sum_{M'\in\mathcal{M}}\exp\left(\frac{1}{2}\text{AIC}(M')\right)}.$$

Note that the scaling of $\frac{1}{2}$ makes the weights proportional to the exponential function of $l_{n,\max} - p$. Thus the weights are proportional to $L_{n,\max}/\exp(p)$, the obtained maximum likelihood divided by e raised to the number of parameters in the model.

A model averaging scheme smoothing the original FIC criterion, has also been developed. In Hjort and Claeskens (2003), the authors suggest to use weights which in our notation may be written as

$$W_{\text{FIC}}(M) = \exp\left(-\kappa\frac{\widehat{\text{FIC}}(M)}{\text{FIC}\left(\widehat{\mu}_{\text{wide}}\right)}\right)\Big/\sum_{M'\in\mathcal{M}}\exp\left(-\kappa\frac{\widehat{\text{FIC}}(M')}{\text{FIC}\left(\widehat{\mu}_{\text{wide}}\right)}\right), \tag{7.4}$$

where κ is a scaling factor deciding how strict the weighting shall be. The factor $\text{FIC}(\widehat{\mu}_{\text{wide}})$ in the scaling is performed in order to make different situations comparable when using the same κ value. Also the negative sign in the exponential is included in order to weight the smallest FIC scores the most, and vice versa.

7.2 Model averaging based on FIC

We will in this section expand the earlier developed FIC schemes to model averaging schemes as introduced above. We will suggest a FIC smoother that weight the μ estimators based on our candidate models according to the obtained FIC scores.

7.2.1 The proposed weight function

The suggested scheme is motivated entirely by the fact that it should be similar to the original FIC smoother. This motivation comes from the fact that the FIC scores are motivated by mse estimation in both situations. We therefore suggests weights on the form:

$$W_{\mathrm{FIC}'}(M) = \exp\left(-\kappa\frac{\mathrm{FIC}(M)}{\mathrm{FIC}_{\mathrm{np}}}\right)\Big/\sum_{M'\in\mathcal{M}}\exp\left(-\kappa\frac{\mathrm{FIC}(M')}{\mathrm{FIC}_{\mathrm{np}}}\right), \tag{7.5}$$

where as usual $\mathrm{FIC}_{\mathrm{np}} = \mathrm{FIC}(\widehat{\mu}_{\mathrm{np}})$ is the FIC score for the nonparametric model. The formulae (7.4) and (7.5) are indeed very similar. The wide model forming the scaling factor in formula (7.4) is replaced by the corresponding model in our setting, i.e. the nonparametric model. This is natural as the nonparametric model may be seen as a parametric model with an infinite number of parameters. In both equation (7.4) and (7.5), the FIC scores are originally estimates of the mean squared error. However, the FIC scores in Hjort and Claeskens (2003) are scaled by a factor $\sqrt{n}$, and subtract a term c which is represented in all FIC scores. The scaling by n does not influence the value of the fraction, but the subtraction of the constant c will cause the κ to correspond to slightly different weighting. Nevertheless, κ values used for smoothing in equation (7.4) should work out well also in equation (7.5).

Note in particular that $\kappa = 0$ corresponds to equal weights for all models independently of the obtained FIC score. As κ increases, more and more weight will be given to the model with the smallest FIC score, which eventually causes the model averaging scheme to weight the winning model by 1 and the others by 0. As a result $\widehat{\mu}_{\mathrm{final}}$ based on this model averaging scheme will coincide with the estimator with the largest FIC score if one simply let κ be large enough.

7.2.2 Limiting distribution and the quiet scandal of statistics

We will now present a lemma providing the precise limiting distribution of certain $\widehat{\mu}_{\mathrm{final}}$ estimators on the form of relation (7.1).

Lemma 7.2.1. *Assume that the following limiting distribution holds:*

$$\sqrt{n}\begin{pmatrix}\widehat{\mu}_{\mathrm{np}}-\mu_{\mathrm{true}}\\ \widehat{\mu}_{\mathrm{pm}^{(1)}}-\mu_{0,\mathrm{pm}^{(1)}}\\ \vdots\\ \widehat{\mu}_{\mathrm{pm}^{(m)}}-\mu_{0,\mathrm{pm}^{(m)}}\end{pmatrix} = \begin{pmatrix}\Lambda_{n,\mathrm{np}}\\ \Lambda_{n,\mathrm{pm}^{(1)}}\\ \vdots\\ \Lambda_{n,\mathrm{pm}^{(m)}}\end{pmatrix} = \Lambda_n \xrightarrow{L} \Lambda = \begin{pmatrix}\Lambda_{\mathrm{np}}\\ \Lambda_{\mathrm{pm}^{(1)}}\\ \vdots\\ \Lambda_{\mathrm{pm}^{(m)}}\end{pmatrix}. \tag{7.6}$$

Let furthermore the final estimator of the focus parameter be given by

$$\widehat{\mu}_{\mathrm{final}} = \sum_{M\in\mathcal{M}} W\left(M;\Lambda_n,\alpha_n\right)\widehat{\mu}_M,$$

for some set of models $\mathcal{M}$, where the weight function has the following properties:

(i) $\sum_{M\in\mathcal{M}} W\left(M;a,b\right) = 1$ *for any* (a,b).

(ii) $W\left(M;a,b\right)$ *is continuous almost everywhere in* (a,b), *for each* $M\in\mathcal{M}$.

(iii) $\alpha_n \xrightarrow{P} \alpha$, *for some constant* α.

(iv) The only randomness of the $W(M;\Lambda_n,\alpha_n)$ *is due to* Λ_n *and* α_n.

Then

$$\sqrt{n}(\widehat{\mu}_{\text{final}} - \mu^*) \xrightarrow{L} \Lambda_{\text{final}} \stackrel{d.}{=} \sum_{M\in\mathcal{M}} W(M;\Lambda,\alpha)\,\Lambda_M,$$

where $\mu^* = \sum_{M\in\mathcal{M}} W(M;\Lambda_n,\alpha_n)\,\mu_{0,M}$, *for* $\mu_{0,\text{np}} = \mu_{\text{true}}$ *and* $\mu_{0,\text{pm}(i)} = \mu_{0,\text{pm}(i)}$.

Proof. Note that by van der Vaart (2000, theorem 2.7 (v)), there is joint convergence in law for Λ_n and α_n, i.e. $(\Lambda_n,\alpha_n) \xrightarrow{L} (\Lambda,\alpha)$. Since the weight function is continuous almost everywhere, $\sqrt{n}(\widehat{\mu}_{\text{final}} - \mu_{\text{true}})$ will also be continuous almost everywhere. The result then follows from the continuous mapping theorem (B.2.9). □

In all situations we will handle, the Λ's are normal. The final limiting distribution appearing in the above lemma is then seen to be a function of several normal distributed random variables. Since the relationship between the variables is not linear in general, the resulting limiting distribution will neither be so. In the rather uninteresting case of nonrandom weights, one will however get a limiting distribution which is normal since the relationship between the random variables then is linear.

The lemma has a nice interpretation, and is clearly general enough to handle different types of schemes presented in this thesis. However, to be fully applicable in a practical situation, it is necessary to add a few restrictions. The following corollary presents an applicable and very fruitful type of weight function under some additional assumption.

Corollary 7.2.2. *Consider the situation of lemma 7.2.1, but where* $\mu_{0,\text{pm}(i)} = \mu_{\text{true}}$ *for all* $i = 1,\ldots,m$. *Consider a FIC scheme consisting of parametric and nonparametric estimators on the following form*

$$\text{FIC}(\widehat{\mu}_M) = \left(\widehat{\text{bias}}(\widehat{\mu}_M)\right)^2 - \widehat{\text{Var}}\left(\widehat{\text{bias}}(\widehat{\mu}_M)\right) + \widehat{\text{Var}}(\widehat{\mu}_M).$$

Let the variance estimators all be $O_p(1/n)$ *and the parametric bias estimators be on the form* $\widehat{\text{bias}}(\widehat{\mu}_M) = \widehat{\mu}_M - \widehat{\mu}_{\text{np}}$. *In addition we assume that* $\text{FIC}_{\text{np}} \neq 0$ *and that* $n\text{FIC}_{\text{np}}$ *is also bounded away from zero with probability 1. Let then* $\widehat{\text{fr}}$ *be a* $(m+1)$*-dimensional vector with elements consisting of the fractions*

$$\widehat{\text{fr}}(M) = \frac{\text{FIC}_M}{\text{FIC}_{\text{np}}},$$

converging in law to $\text{fr}(M)$ *for* $M \in \mathcal{M}$. *Let furthermore* $W'\left(M;\widehat{\text{fr}}\right)$ *be a proper weight function summing to one for* $M \in \mathcal{M}$ *and being continuous almost everywhere in* $c = \widehat{\text{fr}}$. *Then the weight function may be applied to lemma 7.2.1 with* $\mu^* = \mu_{\text{true}}$.

Proof. The proof consists of showing that $\widehat{\text{fr}}$ is a function of Λ_n and α_n only. The rest is immediate from lemma 7.2.1. For the nonparametric model, $\widehat{\text{fr}}(\text{np}) = 1$ which is nonrandom

and obviously not depending on any other random variables. Furthermore, for any parametric model $\text{pm} = \text{pm}^{(i)}$ we have that

$$\begin{aligned}
\widehat{\text{fr}}(\text{pm}) &= \frac{\text{FIC}_{\text{pm}}}{\text{FIC}_{\text{np}}} = \frac{n\text{FIC}_{\text{pm}}}{n\text{FIC}_{\text{np}}} \\
&= \frac{n(\widehat{\mu}_{\text{pm}} - \widehat{\mu}_{\text{np}})^2 + O_p(1)}{O_p(1)} \\
&= \frac{\left(\sqrt{n}(\widehat{\mu}_{\text{pm}} - \mu_{\text{true}}) - \sqrt{n}(\widehat{\mu}_{\text{np}} - \mu_{\text{true}})\right)^2 + O_p(1)}{O_p(1)} \\
&= \frac{(\Lambda_{n,\text{pm}} - \Lambda_{n,\text{np}})^2 + O_p(1)}{O_p(1)}. \qquad (7.7)
\end{aligned}$$

Now, since the expression in the denominator is by assumption bounded away from zero, expression (7.7) is a valid also in the limit. Consequently, $\widehat{\text{fr}}(\text{pm})$ is seen to depend only on Λ_n and terms converging in probability to constants. The result then follows from lemma 7.2.1. □

The situation in the above corollary holds for the schemes of chapter 3 since we have shown that the these variance estimators are consistent. Although, for the limiting distribution to hold, we must assume that each of the parametric distributions fulfills $\mu_{0,\text{pm}} = \mu_{\text{true}}$. This is particularly the case whenever all of the parametric models are true or they are only locally misspecified. The latter situation occurs e.g. when all parametric models are special cases of the model with the most parameters, as introduced in Claeskens and Hjort (2008, chapter 5), or more generally the situation handled in appendix A. Furthermore, the same applies to the situations of censored data handled in chapter 4, provided that the variance estimators are consistent.

Note however that we have only given joint limiting distributions for one nonparametric and one parametric distribution at a time for the situations mentioned above.[2] However, the derivations leading to the corresponding limiting distributions in chapter 3, chapter 4 and appendix A, are all based on central limit theorems. Thus, the generalization to joint convergence of the nonparametric estimator and possibly q different parametric estimators are immediate, whenever all the parametric models meet the stated conditions.

Corollary 7.2.2 is not interesting without a weight function. As the careful reader may already have noticed, the weight function of equation (7.5) is fortunately on the required form, since we may write

$$W_{\text{FIC}}(M) = W'_{\text{FIC}}\left(M; \widehat{\text{fr}}(M)\right).$$

Therefore, when $\mu_{0,\text{pm}} = \mu_{\text{true}}$ for all the parametric distributions, the several FIC schemes may be formed into model averaging schemes with weights as given by expression (7.5). Those schemes then have the property that

$$\sqrt{n}(\widehat{\mu}_{\text{final}} - \mu_{\text{true}}) \xrightarrow{L} \sum_{M \in \mathcal{M}} W'_{\text{FIC}}\left(M; \text{fr}(M)\right) \Lambda_M. \qquad (7.8)$$

As discussed when introducing this model averaging scheme, letting $\kappa \to \infty$ corresponds to basing all further inference solely on the FIC winning estimator. Provided that taking the limit

[2] This was done as joint limits of all parametric estimators was not needed in any previous arguments, and it is clearly easier to read a presentation where one does not repeat every argument for q different parametric models.

as $\kappa \to \infty$ on both sides of result (7.8) above is allowed, then the limiting distribution of the estimator chosen by FIC is directly obtained. A condition for this to be valid is somewhat loosely that the maximum distance between the limiting distribution for finite and infinite κ reduces to zero when $\kappa \to \infty$. For further details and a more precise condition, see e.g. Billingsley (1999, chapter 1). In any case the limiting distribution for the FIC-winning distribution may be derived by representing the estimator based on FIC as in relation (7.2), and applying lemma 7.2.1 with that weight function.

Obtaining the limiting distribution of the final estimator based on FIC may be quite useful. The limiting distribution of the "post-selection" estimator $\widehat{\mu}_{\text{final}}$ is the limiting distribution of the chosen estimator when the model selection step is taken into account. Model selection is used much more in newer times, than was the case before computers made it possible to fit statistical models just by a few punches on a computer. This is clearly a good thing – making inference more reliable than just guessing that some model is the best we can do. However, most often the inference and work after a model has been selected is carried out without taking the model selection step into account. Inference is done using the selected model without even considering the fact that another model could in fact have been chosen. Since the data are random, another data set stemming from the exact same distribution could have resulted in a totally different model being selected, and totally different uncertainty estimates, confidence intervals etc. Consequently, the uncertainty estimates and confidence intervals obtained when ignoring the additional randomness, tend to be too optimistic and narrow. This possible flaw which is performed by many researchers and even trained statisticians, is known as the "quiet scandal of statistics". As pointed out above, the limiting distribution of the final estimator is often multimodal with clear distinctions from the normal distributions which inference often is based on. Techniques similar to those of Hjort and Claeskens (2003) may furthermore be used to calculate the actual limiting coverage probability of confidence intervals used when ignoring the model selection step. In addition the authors suggest modified confidence intervals to adjust for this additional uncertainty. Similar techniques may be applied for our situations, but for now we are content by pointing out this important point. We also mention that calculating uncertainty estimates under the other competing models, especially if the winning model was not a clear winner, may help to indicate in what direction the confidence interval is to optimistic. More on the quiet scandal of statistics and repaired confidence intervals are given in Claeskens and Hjort (2008) and Buckland et al. (1997).

7.2.3 Limiting distribution under parametric truth

The above methodology is a bit complicated since it is stated in quite general terms. For the sake of illustration we will therefore write out the limiting distribution for the situation where there is only one parametric model and that model is also fully correct. We concentrate on the regular iid situation of chapter 3.

For this particular situation, we have earlier shown that the nonparametric model is selected whenever

$$n\widehat{b}^2 \geq 2(\widehat{V}_{\text{np}} - \widehat{V}_{\text{pm,np}}).$$

Hence, the winning estimator based on this scheme may be written

$$\widehat{\mu}_{\text{final}} = L_n(\sqrt{n}\widehat{b})\widehat{\mu}_{\text{np}} + (1 - L_n(\sqrt{n}\widehat{b}))\widehat{\mu}_{\text{pm}},$$

where $L_n(z) = \mathbf{1}_{\{z^2 \geq 2(\widehat{V}_{\rm np} - \widehat{V}_{\rm pm,np})\}}(z)$. Consequently, we have that

$$\sqrt{n}(\widehat{\mu}_{\rm final} - \mu_{\rm true}) = L_n(\sqrt{n}\widehat{b})\sqrt{n}(\widehat{\mu}_{\rm np} - \mu_{\rm true}) + (1 - L_n(\sqrt{n}\widehat{b}))\sqrt{n}(\widehat{\mu}_{\rm pm} - \mu_{\rm true}). \tag{7.9}$$

Furthermore, by writing

$$\Lambda_{n,\rm np} = \sqrt{n}(\widehat{\mu}_{\rm np} - \mu_{\rm true}), \qquad \Lambda_{n,\rm pm} = \sqrt{n}(\widehat{\mu}_{\rm pm} - \mu_{\rm true}),$$

which also gives

$$\sqrt{n}\widehat{b} = \Lambda_{n,\rm pm} - \Lambda_{n,\rm np},$$

we may rewrite equation (7.9) as

$$\sqrt{n}(\widehat{\mu}_{\rm final} - \mu_{\rm true}) = L_n\left(\Lambda_{n,\rm pm} - \Lambda_{n,\rm np}\right)\Lambda_{n,\rm np} + \left(1 - L_n\left(\Lambda_{n,\rm pm} - \Lambda_{n,\rm np}\right)\right)\Lambda_{n,\rm pm}.$$

Under the regularity conditions of lemma 3.5.1, we have found that

$$\begin{pmatrix} \Lambda_{n,\rm np} \\ \Lambda_{n,\rm pm} \end{pmatrix} \xrightarrow{L} \begin{pmatrix} \Lambda_{\rm np} \\ \Lambda_{\rm pm} \end{pmatrix},$$

where $\Lambda_{\rm np} \sim N(0, V_{\rm np})$ and $\Lambda_{\rm pm} \sim N(0, V_{\rm pm})$. In addition $\mathrm{Cov}(\Lambda_{\rm np}, \Lambda_{\rm pm}) = V_{\rm pm}$. Using the continuous mapping theorem (B.2.9), we get the following limiting distribution

$$\sqrt{n}(\mu_{\rm final} - \mu_{\rm true}) \xrightarrow{L} L(\Lambda_{\rm pm} - \Lambda_{\rm np})\Lambda_{\rm pm} + (1 - L(\Lambda_{\rm pm} - \Lambda_{\rm np}))\Lambda_{\rm np},$$

where $L(z) = \mathbf{1}_{\{z^2 \geq 2(V_{\rm np} - V_{\rm pm})\}}(z)$. This means that the resulting limiting distribution of the final estimator based on FIC is a nonlinear combination of two normal distributed variables. The limiting distribution has a fairly easy form and is also straight forward to sample from. The resulting distribution may however be highly non-normal and even bimodal.

7.3 An example of model averaging in use

In this example we will once again consider the data of the Norwegian population's activity level as treated both in an example in section 3.9.1 of chapter 3 and then again in section 6.3 of the previous chapter. For a review of these data, we refer to the former section.

Here we will simply apply the model averaging scheme stated in equation (7.5) when considering the same models and focus parameter as in the original example. Recall that the data consist of the daily number of steps, and that we fitted the normal, the log-normal and the skewed normal distribution in addition to the usual nonparametric model. The focus parameter was $\mu(H) = 1 - H(10000)$.

To perform model averaging on this situation, we must specify the value of κ in formula (7.5). For the sake of illustration, let us apply the model averaging scheme with the three different scales $\kappa = 0.5, 1, 6$ corresponding to mild, quite balanced and hard weighting. The returned weights and final estimators are given in table 7.1.

Recalling that the skewed model was declared the winner based on FIC, it appears from the table that the final estimates based on these model averaging schemes are smaller than was the case when we only relied on one model. The background for this lies in the fact that the second best model is the nonparametric model, which has a smaller estimate. We also observe the changes both in the weights and the resulting estimate for increasing κ. For mild

		W_{FIC}		
Model	$\widehat{\mu}$	$\kappa = 0.5$	$\kappa = 1$	$\kappa = 6$
Nonpar	0.2438	0.3678	0.3645	0.0808
Normal	0.2776	0.002	0.000	0.000
Skewed normal	0.2494	0.4504	0.5466	0.9192
Log-normal	0.2562	0.1816	0.0899	0.000
$\widehat{\mu}_{\text{final}}$		0.2486	0.2479	0.2489

Table 7.1: Weights and final estimates for model averaging with different FIC weights using formula (7.5).

weighting all models possibly except the normal model influence the final estimator, for the quite balanced weighting it is mostly the two best models that has major influence, whereas the hard weighting gives almost all weight to the winning model. Note that the normal model is such a poor choice that even with mild weighting, the estimate based on the normal model has negligible contribution.

Which value one should use for κ would vary from situation to situation depending on how much one wishes to trust the winning model. It is difficult to give any advice on a good choice of κ for a general situation. It does not seem like a κ value around 1 is the worst choice, but to make a well qualified choices of κ more experience with the machinery is beneficial.

Chapter 8

FIC in R

Having proposed new model selection criteria, it is certainly beneficial if they can be applied to practical problems without too much trouble. Along with the thesis an `R` function has been developed to perform automatic FIC analysis for the iid situation in the standard framework of chapter 3. This chapter presents and attempts to explain how the function is used without having to bore the reader by going into the details on how all computational details are solved. Still, making the function flexible enough to handle a great number of situations has caused it to have quite many input arguments. To get started with the function it may thus be helpful to get an explanation of each of the input variables. Therefore, we start this chapter by a short explanation on how the function is built and working. Then we continue by explaining how and when the different input variables should be used, before we finish off by discussing and presenting a few short and simple scripts on how the function in use.

The source code for the function may be found at `http://folk.uio.no/martinju/FIC`. On the same web page all `R` scripts from the examples of this thesis are provided, in addition to the source code of a simple version of the wFIC scheme for iid data with discrete weights that also has been developed.

8.1 Development and structure of the program

Although the equations involved in the main FIC scheme does not seem very complicated, developing this FIC function has demanded great effort. The main reason for the great amount of time that has been spent to create this `R` function, is that it has been a goal to make it flexible, easy to use, precise and fast, all at the same time. Hopefully this goal has also been achieved. The function is flexible in the way that it can handle any parametric model which can be precisely described by a standard density curve or pmf plot. Furthermore, it is flexible in the way that when exact analytical expressions exists for variables involved in the calculations, these can be used, but in the many cases where analytically expressions do not exist (or when the user does not bother to derive them), numerical approximations with high precision are automatically used without any extra effort from the user. Some of the numerical algorithms are discussed further down in this chapter. There are a number of preset focus parameters (included all the most common ones) that the user can apply just by specifying the name corresponding to it. In addition the user may specify any focus parameter of the form of smooth functions of averages, discussed in this thesis. The user only specifies the ϕ and S

functions in

$$\mu(H) = \phi\left(\int S(x)\,\mathrm{d}H(x)\right).$$

The function should furthermore be fairly easy to use and a number of different help, error and warning messages have been included to help the user to specify the right input variables leading to correct FIC analysis. The function consists of almost 1500 lines of code. Most of the code is just definitions and checks that everything is working as it should, nonetheless some computations takes time. To reduce the time of computations, vectorized objects are used all the way rather than looping – in spirit of good programming. However, when it comes to numerical approximations, speed and precision does not go hand in hand. Getting precise results has been given the highest priority here to make sure that the produced results are reliable. In most cases, the extra time needed to be more precise, is anyway a matter of seconds, not minutes or hours. Finally, the function is tested on a wide range of problems, a great number of bugs are fixed during the development, and the current version appears stable. Nonetheless, bugs may still appear and there is no guarantee of the produced results.

8.2 The function and its input variables

We are now going to explain the input arguments of the function and how it operates. For a more comprehensive and detailed walkthrough, we refer to the actual script code with comments. In `R` the function appears as below:

FICfunc

```
FICfunc=function(Data,data.type="cont",FICmethod="standard",numerical=NULL,focus.preset="mean",
focus.preset.extra=NULL,focus.user=NULL,parmodels.preset="gaussian",parmodels.user=NULL,
MLestimates=NULL,print.res=TRUE,print.warnings=TRUE,make.plot=TRUE)
```

The great number of input variables is due to the flexibility of the function. Note however that in the simplest and most straight forward cases, just a few of these arguments needs to be specified, since the default values of them produces the desired type of analysis. The main output of the function is a nice looking table with the key values of the scheme. In addition, warning messages pointing out possible strange behavior of results (such as negative squared bias estimate), and a plot of a histogram and the fitted density curves of pmf plots of the parametric distributions are also provided. Also, everything that is calculated in the function and all input variables are stored in objects the user can investigate further if interested. The functions running time before giving the output are also timed and saved in an object.

When loading the function code, the code checks if the package `numDeriv` is installed and loaded, if it is not this is done. The package contains functions performing numerical derivation using well-tested and precise methods. The first part of the actual FIC function contains a number of `if` tests to check that all of the input arguments are given on the correct form. In the cases where they are not on the correct form or not specified at all, the function stops and gives a help message to help the user specify it correctly. Then the function checks if the combination of the input arguments and objects match each other. The next part of the FIC function contains definitions of the functions that are used in the actual calculations later on. The calculations are then performed for the nonparametric and all the parametric models according to how the user has specified that this should be done. Finally, these results are stored in convenient objects, and results are printed together with possible warning messages and the plot of the histogram and the density or pmf of fitted parametric models.

We now turn to the input arguments of the FIC function. The arguments of the function are explained in the following list:

- `Data`: A required numeric argument of any positive length. This is the data set to be evaluated by the function. Any numbers can be given to it (NA and nan are removed)
- `data.type`: A character of length 1. This argument decides whether the data should be treated as continuous or discrete. The parametric models fitted must be of the same type. Either `"cont"` (default) or `"disc"` can be inserted. The former corresponds to continuous data and the latter to discrete data.
- `FICmethod`: A character of length 1. This decides which FIC formula that should be used in this setting. Either `"standard"` (default) or `"standard_adj"` can be inserted. The former gives the main scheme, whereas the latter gives the version adjusting for possible negative estimates of squared bias and variance.
- `numerical`: An vector of characters of any length or `NULL`. The object specifies what part of the calculation that should be done numerical. If this is set to `NULL`, all calculations are done exact (as far as possible). If the vector includes
  - `"deriv"`: The derivation of the focus parameter with respect to the parameters of the parametric models are done numerically.
  - `"score"`: The score function is calculated numerically.
  - `"info"`: The information function is calculated numerically.
- `focus.preset`: A character of length 1 or `NULL`. This argument specifies which (if any) of the preset focus parameters that shall be used. The user can choose any of the following preset focus parameters `"median"`, `"quantile"`, `"mean"`, `"var"`, `"sd"`, `"cumulative"`, `"madam"` , `"IQR"` and `"prob.mass"`.
- `focus.preset.extra`: A numeric of length 1 or `NULL`. The argument is used to specify an additional argument to the preset focus parameter (if necessary). Especially it is required when `focus.preset` is set to `"quantile"` `"cumulative"` or `"prob.mass"`, to specify which point of that is of interest. The argument is not used for other focus parameters.
- `focus.user` A list or `NULL`. The argument specifies any so-called smooth focus parameter $\mu(H) = \phi\left(\int S(x)\,\mathrm{d}H(x)\right)$, of the user's choice. `NULL` is default. The list must contain the following arguments:
  - `name`: A character of length 1 giving the name of the focus parameter,
  - `s`: A function of 1 argument specifying the S function.
  - `phi`: A function of 1 argument specifying the ϕ function.
- `parmodels.preset`: A vector of characters of any length or `NULL`. The argument specifies which (if any) of the preset parametric models that should be fitted. By default `"gaussian"` is included in the model selection scheme when `data.type` is set to `"cont"`, and none is included when it is set to `"disc"`. As of September 2012, this distribution is also the only preset distribution that is included.

- `MLestimates`: A list of vectors or `NULL`. The argument specifies the fitted ML estimates of the parametric models. This is particularly useful when fitting a parametric model where analytic expressions for the ML estimators does not exist, e.g. for the gamma or weibull distributions. In those cases an `nlm` algorithm may be applied first to find the ML estimates numerically before this FICfunc is used – more on this in the next section.

- `parmodels.user`: A list of lists or `NULL`. The argument specifies any parametric model of the user's choice. `NULL` is the default, meaning that no parametric model other than the preset one, is included. To specify parametric models, this list must contain a new list for each new parametric model. The name of the new list is the name of the model, and it must contain arguments depending on how the focus parameter is specified or which parts of the calculation that should be solved numerically. The following objects are always required:

  - `dim`: A numeric of length 1 specifying the dimension of the parametric model.
  - `pdf`: A function of the evaluation point and the parameter values (2 arguments) returning the probability distribution.
  - `log.pdf`: The same as `pdf` but the logarithm of the function value is returned instead.

  In addition the following objects are required depending on other input variables:

  - `cdf`: A function the evaluation point and the parameter values (2 arguments) returning the cdf, is required for any non-smooth focus parameter.
  - `ML`: A function of the data (1 argument) returning the ML estimates is required when `MLestimates` does not include a vector corresponding to this particular parametric model.
  - `deriv$(name of the focus parameter)`: A function of the parameter value (1 argument) returning the derivative of the focus parameter with respect to the parameters of the parametric model, is required if not `"deriv"` is included in `numerical`.
  - `score`: A function of the data point value and the parameter values (2 arguments) returning the score function of the parametric model, is required if not `"score"` is included in `numerical`.
  - `info`: A function of the data point value and the parameter values (2 arguments) returning the information function of the parametric model, is required if not `"info"` is included in `numerical`.
  - `eval. set`: A numeric of any positive length specifying the evaluation points for a discrete distribution. Required if `data.type="disc"`.
  - `pdf`: A function of the eval.set and the parameter values (2 arguments) returning the probability mass for a discrete distribution. Required if `data.type="disc"`.
  - `qdf`: A function of the eval.set and parameter values (2 arguments) returning quantile function for a discrete distribution function. Required if `data.type="disc"` and `focus.type!="smooth"`.

- `print.res`: A logical argument (`TRUE` or `FALSE`). The argument specifies if the results should be printed to the terminal or not. This is useful if `FICfunc` are used repeatedly for

simulation studies or similar, where the result table of every execution is not of interest. The default is `TRUE`.

- `print.warnings`: A logical argument. The argument specifies if warnings should be printed or not. This can also be useful in simulation studies or whenever unnecessary printing should be omitted. The default is `TRUE`.
- `make.plot`: A logical object. It specifies if a histogram with plotted density curves of the parametric distribution should be plotted or not. The default is `TRUE`.

Note that exactly one focus parameter must be specified. If several focus parameters are of interest, simply run the function again with another focus parameter. Also, either fitted ML estimates must be provided through `MLestimates` or a function producing these values from the data must be provided by `parmodels.user$ML`. In addition to printing a table with the summary results from running the scheme, the function also stores many of the intermediate results obtained while running. This may be particularly useful to validate the results or to investigate what caused the particular output. Also, when using the function for simulation purposes this is useful.

8.2.1 Numerical algorithms

As mentioned already in section 3.2, some situations are not solvable without numerical approximations. This concerns mainly quantities which are based on differentiation or integration. Since accuracy is important in this situation, we do not try to build something clever on our own, but rely on built-in algorithms in the statistical programming language `R` instead, implemented by people with much better knowledge of this. For numerical differentiation, we use the functions `grad` and `jacobian` of the package `numDeriv` and the method `"Richardson"`, which uses so-called Richardson extrapolation. Although Richardson extrapolation is not the fastest method available, it is known to be of the most precise and reliable methods, which is why we choose to use exactly this method.

Numerical integration using the function `integrate` in `R` performs one-dimensional numerical integration based on the so-called "quadpack" routine, see Piessens et al. (1983). This works out fairly well by placing an additional simple self-made algorithm, wisely choosing the integration bounds in each situations. More on these numerical derivation and integration techniques can be found in any book on numerical analysis

8.3 The program in use

We finish off this short chapter by providing a few examples of `FICfunc` in use. We start out with possibly the simplest meaningful example one could think of: Testing whether a normal distribution or the nonparametric model is best at estimating the median of a data set. The test data are provided by simulating a set of 100 variables from the standard normal distribution. To perform this task, we run the following simple code in `R`:

Command line arguments

```
source("http://folk.uio.no/martinju/FIC/source_code/FICfunc.final.final.R") # Loading the FIC function

# The simplest test ever
set.seed(321)     # Setting some seed level to ease reproduction.
data=rnorm(100)   # Simulate the data
```

```
Test=FICfunc(data,focus.preset="median")        # Running the FIC function
```

When executed `FICfunc` prints the following output:

Output

```
> Test=FICfunc(data,focus.preset="median")
                   mu dim      bias*         sd       RMSE    Rank
Nonpar   0.086579255 Inf 0.00000000 0.11224232 0.11224232       2
Gaussian 0.009067669   2 0.03029794 0.09509784 0.09980764       1

 Warning: To estimate the influence values of the median, a non-plug-in estimation procedure has been
                used to estimate the density at the focus parameter g(mu). The function "density" with
                the default bandwidth in R is used to do this.
```

We see a nice and hopefully intuitive table is printed. On the far left the model name is given, furthermore, `mu` denotes the estimates of the focus parameter, `dim` the number of parameters in the model (where the nonparametric is always set to 0). Then the key estimation results of squared bias and variance are brought back to scale by taking the root of the estimates. The reason we do this is that it is easier to compare the sizes of them when they are on the same scale as μ. Thus, `bias*` is the square root of the absolute value of the squared value estimate and `sd` is simply the square root of the variance estimates. Moreover, `RMSE` denotes the square root of the FIC scores. Once again the square root is taken to make it easier to compare the sizes of the terms. Finally, the last column gives the ranking of the models. Below the table a warning message has been given, since an estimation method different from the usual plug-in technique has been used. When the adjusted scheme is used instead, a subscript "adj" is added to the terms `bias* sd` and `RMSE` to mark that the formulae are adjusted.

We are now going to show how a slightly more complicated model selection task may be handled by this function. We will use the example given in section 3.9.1 concerning the Norwegian population's activity level. For this situations we have data on the number of daily steps for each person and are interested in the proportion of the population that walks more than 10 000 steps. Thus the focus parameter is $\mu(H) = 1 - H(10000)$ for a cdf H. This focus parameter is not available as one of the preset focus parameters, so we must manually specify it first. However, the focus parameter is clearly on the smooth form and specified by $S(y) = \mathbf{1}_{\{y \geq 10000\}}(y)$ and $\phi(y) = y$. In `R` we define the following functions to deal with this:

Focus parameter definition

```
focus.user=list()
focus.user$name="Proportion"
focus.user$s=function(val)
        {
        (val>=10000)
        }

focus.user$phi=function(val)
        {
        val
        }
```

Furthermore, we included three different parametric model in the set of candidate models, where only the Gaussian model is preset in the FIC function. Thus, we have to do just a bit of work first to apply the two other parametric distribution. Firstly the lognormal distribution is to be added. Since the ML estimator for this parametric distribution can be found analytically, it is preferred to calculate them and include them as formulae. The following script defines the

list of functions needed for the lognormal distribution to be included in the set of candidate models:

Adding the lognormal distribution

```
# Adding the lognormal distribution

parmodels.user.list=list()
parmodels.user.list$lognormal$dim=2
parmodels.user.list$lognormal$pdf=function(val,para)
                {
                (val>0)*dlnorm(val,para[1],para[2])
                }

parmodels.user.list$lognormal$log.pdf=function(val,para)
                {
                (val>0)*dlnorm(val,para[1],para[2],log=TRUE)
                }

parmodels.user.list$lognormal$ML=function(data)
                {
                c(mean(log(data)),sd(log(data)))
                }
```

For the skewed normal distribution, however, there exists no analytical formula for the ML estimator. Therefore, the `nlm` function which performs nonlinear minimization, is applied to find the maximum of the log-likelihood function. Some trial and failure might be needed concerning starting values of the numerical algorithm. In this example it is however pretty straight forward. Before we can run this algorithm we define the necessary function of the skewed normal distribution by importing the package `sn`.

Adding the skewed normal distribution

```
# Includes the skewed normal distribution

#install.packages("sn")
library("sn")                           # Installs the skew normal distribution.
parmodels.user.list$sn$dim=3

parmodels.user.list$sn$pdf=function(val,para)
                {
                dsn(val,dp=para)
                }

parmodels.user.list$sn$log.pdf=function(val,para)
                {
                dsn(val,dp=para,log=TRUE)
                }

# Estimates the parametric parameters with the nlm function
minloglik.sn=function(para)
        {
        -sum(parmodels.user.list$sn$log.pdf(data,para))
        }
nlm.sn=nlm(minloglik.sn,c(mean(data)+100,sd(data)+100,shape=1),gradtol=10^(-15),steptol=10^(-15))
nlm.sn$code           # Printing the code to check that the nlm function has converged correctly.

MLest=list()
MLest$sn=nlm.sn$estimate
```

Now, as the focus parameter is defined along with all the necessary definitions for the parametric models, we can apply the FIC function to this problem by inserting arguments corresponding to the FIC analysis we wish to run.

Executing FICfunc for example 1

```
FIC_steps=FICfunc(Data=data,FICmethod="standard",numerical=c("deriv","score","info"),focus.preset=NULL,
focus.user=focus.user,parmodels.preset="Gaussian",parmodels.user=parmodels.user.list,MLestimates=MLest)
```

Note that both derivation with respect to the parameters in the parametric models and finding the score and information functions are done numerically in this situation. For the lognormal distribution we could have found analytical expressions for these quantities, but we expect it to be hard for the skewed normal distribution, and therefore decided to use numerical algorithms instead. The use of numerical algorithms do however increase the execution time of the function, but even if the amount of data are not small here, it all takes less than half a minute on a lightweight modern laptop. As seen in the above scripts, some more work has to be done in advance of executing the FIC function, and some more arguments needs to be specified compared to the first simple situation of the median. Still we believe this is not overwhelming for most statistical researchers.

Chapter 9

Concluding remarks

To summarize we give a short review of what has been achieved through the thesis. We also discuss and make a few concluding remarks upon the main contributions of the thesis. Finally we discuss a few topics for further research.

Summary and discussion

In this thesis we developed focused information criteria (FIC) for a number of situations concerning nonparametric vs. parametric model selection, when focus was on the parameter μ. We mainly concentrated on fully observed iid data, but have also developed FIC schemes for other data types, including censored iid data and the situation of two samples. All the different FIC schemes developed here were in some way or another based on estimates of the mean squared error of μ estimators under each of the models. The model with the smallest estimate was consequently selected. Most of the developed schemes were based on the asymptotic behavior of the μ estimators and consisted of estimators of the two terms of the mean squared error: the squared bias and the variance. The most frequent strategy was to first derive joint limiting distributions for each pair of nonparametric and parametric estimators, approximate the mse for each of the μ estimators based on these, and finally estimate the involved quantities using the data.

The vast part of the thesis concerned fully observed iid data, where we also feel the main contribution of the thesis lies. For the most regular situation where the nonparametric estimator is simply the plug-in estimator $\mu(\widehat{G}_n)$, a number of properties of the derived FIC scheme where carefully investigated. For instance, rather weak sufficient conditions were stated under which the key result and the scheme is fully working. Furthermore, strong consistency for the variance estimators in the scheme were obtained along with similar results for the squared bias estimators. This is a strong property of the apparatus, which also made it possible to investigate the asymptotical properties of the scheme in a convenient manner. Consequently we showed that the scheme may be seen as an implicit hypothesis test with an asymptotic level of 0.157. In contrast to other information criteria choosing among only parametric models, it was shown that the FIC scheme select the best model with probability 1 whenever $\mu_{\text{true}} \neq \mu_{0,\text{pm}}$ for all the parametric models. By excluding some special cases, one may state this as: When the sample size grows, the main FIC scheme tends to perform better than any information criterion selecting between only parametric models whenever none of the parametric models are exact. This is clearly a very strong property. Performance studies also indicated that the FIC apparatus is competitive with AIC and BIC for moderate sample sizes, even when one of

the parametric models actually are fully correct.

In addition to the careful investigations for situations of fully observed data, we handled and discussed censored iid data quite a bit. The schemes were however restricted to the most common cases and were thus not as generally stated as their precursor. Neither were the implications investigated in such great detail. The approach did however turn out fruitfully, and opened the world to a whole lot of new situations. We also attempted to derive FIC formulae for a few situations not handled by the most general schemes. For instance, we succeeded in creating useable formulae for density estimation and schemes for some types of focus parameters based on two samples. For the more general setting of regression models, we discussed and showed why the expansion is difficult but still indicated how one may carry out focused model selection also in this setting.

In addition to criteria focusing on one single focus parameter, we presented a generalization which incorporates several focus parameters at the same time. This was done in a general fashion only depending on an existing FIC scheme of the type we previously had discussed. We further showed that the regular iid version of this scheme (wFIC) had close connections to a certain class of goodness of fit tests, and also indicated parameter independence for such a wide class of distributions. For testing normality we obtained an asymptotic level which were surprisingly close the usual level of 0.05 which somewhat artificially often are chosen among statisticians. This was in particular a connection we were glad to find.

The theme of model averaging was also touched. A general model averaging apparatus with the usual FIC scheme as a special case was introduced. We also derived the limiting distribution of the final estimator based on this general apparatus under some additional assumptions. In addition we pointed out that the limiting distribution is in general non-normal, despite what one would imagine. This last point is certainly of great importance when basing further inference on the final estimator.

Further work

Although we were able to reach quite a few results through this thesis, there are still many unexplored paths to follow.

One of the first things we would have explored in greater detail if the time frame allowed it, is the connection between wFIC and a certain goodness of fit test. This was one of the last themes we studied, and even though we managed to show the connection between the two themes fairly well, there are still more work to be done on this theme. As mentioned, it should be possible to derive the limiting distributions analytically via process theory and uniform convergence. We simply did not have time to tackle this via this superior approach and therefore followed the simulation approach instead. The same applies to the parameter independence. Some more work would have been necessary to show parameter independence for a wider class. However, the approach and results we managed to give were of such great interest that we decided to include them even if a more careful treatment of these themes would have been desirable.

If the time frame of the thesis had made it possible, it would have been interesting to also investigate some of the following ideas:

- Although the regression situation was discussed in the thesis, we did not spend too much time trying to derive precise FIC formulae in that framework. Lifting the scheme to such

a setting and giving precise FIC formulae would certainly be of great practical interest. The amount of work needed to follow this idea to an end is however surely quite large.

- When the general regression setting discussed above seems rather difficult, a regression setting for censored data may seem more accessible. Even though the following situation is more of a semiparametric vs. parametric theme rather than the usual nonparametric vs. parametric, the natural strategy has clear similarities to the strategies discussed in this thesis. Consider the Cox proportional hazard model where the hazard function $\alpha(t|x)$ for a covariate vector x may be written as $\alpha(t|x) = \alpha(t)\exp(x^t\beta)$, for some parameter vector β. Considering e.g. the situation when the focus parameter is the cumulative hazard rate at time t for an individual with covariate vector x, the nonparametric and parametric estimators may be represented as follows:

$$\widehat{\mu}_{\rm np}(x) = \widehat{A}_{\rm np}(t)\exp(x^t\widehat{\beta}_{\rm np}),$$
$$\widehat{\mu}_{\rm np}(x) = \widehat{A}_{\rm pm}(t,\widehat{\theta}_n)\exp(x^t\widehat{\beta}_{\rm pm}).$$

  Here, the regression factor for the nonparametric estimator may be estimated by Cox regression, and the cumulative hazard factor may be estimated by the Breslow estimator, see e.g. Andersen et al. (1993). The parametric estimators may be estimated by maximum likelihood techniques as discussed in Hjort (1990, section 6). The reason for this being a more attainable approach than other regression settings, is that all estimators converge with regular $\sqrt{n}$ rate, which were seen not to be the case in the situation studied in section 5.2.

- All our derivations in this thesis have been based on first order asymptotics. However, as indicated in section 3.6, it may seem like FIC does not estimate the mse that well when the focus parameters are quite complicated. One possible reason for this is that the first order approximation is not precise enough when the focus parameter is complex. Higher order approximations e.g. through higher order Taylor expansion may in such a context be a natural starting point.

- We have throughout the entire thesis focused on selecting the model whose estimator has the smallest mean squared error. In other words we have been aiming at minimizing the risk function given by the expected loss for the loss function

$$L(\widehat{\mu}) = (\widehat{\mu} - \mu_{\rm true})^2.$$

  This is for sure the most studied loss function, both because it is intuitive and self-explanatory, but also since the risk function behaves so nicely by splitting into squared bias and variance. However, for some situation this loss function may not be as fruitful as other loss function. The absolute loss $L(\widehat{\mu}) = |\widehat{\mu} - \mu_{\rm true}|$ is for sure a nice loss function where one is penalized harder for being slightly wrong and not as hard for being very wrong, compared to the squared loss function. Another loss function that penalizes more for underestimation than overestimation or vice versa depending on some scalar a, is the so-called Linex loss invented by Varian (1974). The loss function may be written as

$$L(\widehat{\mu};a) = \exp(a\Delta) - a\Delta - 1,$$

  where $\Delta = \frac{\widehat{\mu}-\mu_{\rm true}}{\mu_{\rm true}}$. This loss function is certainly of great importance in econometry where overestimating a profit or at least underestimating a loss, is of greater danger

than the other way around. The Linex loss behaves rather nicely in terms of being well approximated by a Taylor expansion for small values of a. This behavior creates an excellent starting point for asymptotic analysis.

- Finally, it would of interesting to expand the situation of comparing two samples from section 5.3 to more general comparison functions. Also comparisons of more than two samples at the same time would increase the scheme's usefulness.

Appendix A

Limit results in a locally misspecified framework

In this appendix we derive limit results based on a locally misspecified framework. The limit results are used in the sections 3.4.4 and 3.5.2 of chapter 3 to investigate respectively why the squared bias estimator should be on the chosen form and a more informative the limiting selection probability in such a framework. In section 5.4 we used the result of this appendix to develop a FIC scheme directly.

In this framework we will assume that iid data $Y_1, \ldots, Y_n$ stems from a true distribution that varies with the sample size n.[1] Especially we assume that the density (for continuous distributions) or probability mass function (for discrete distributions) takes the form

$$g_n(y) = f(y; \theta_0) + \frac{r(y)}{\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right), \tag{A.1}$$

where $r(y) : \mathbb{R} \to \mathbb{R}$ is assumed to be a function independent of the sample size n and not necessarily continuous, but with the property that $\int r(y)\,\mathrm{d}v(y) = 0$. We also assume that any integral including the $o(1/\sqrt{n})$ term, where the rest of the integrand is of size $O(1)$, gives $o(1/\sqrt{n})$. As usual f_θ is the density (or probability mass function) of a parametric distribution. Consequently, the cdf of this distribution may be written as

$$G_n(y) = F(y; \theta_0) + \frac{R(y)}{\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right), \tag{A.2}$$

for $R(y) = \int_{-\infty}^{y} r(z)\,\mathrm{d}v(z)$. This framework is locally misspecified in the sense that g_n is a distance $O(1/\sqrt{n})$ away from the parametric model evaluated at the least false parameter θ_0. This framework is a generalization of the framework used in Claeskens and Hjort (2003) and Hjort and Claeskens (2003) to obtain the classical FIC and evaluate properties of certain model averaging schemes. These authors work with the framework where

$$g_n(y) = f(y; \theta_0, \gamma_0 + \delta/\sqrt{n}), \tag{A.3}$$

[1] Note that there is a slight abuse of notation here since we do not emphasize that the data depends on n. To be fully precise we should have written $Y_{n1}, \ldots, Y_{nn}$, but for convenience and comparison with the other notation, this is dropped here.

is the true distribution ,and γ_0 corresponds to the value of an additional parameter γ which reduces the wider model $f(y;\theta,\gamma)$ to the narrower model $f(y;\theta)$. A Taylor expansion of this function does gives

$$g_n(y) = f(y;\theta_0,\gamma_0) + U'(y;\theta_0,\gamma_0)\delta/\sqrt{n} + o(1/\sqrt{n}),$$

where $U'(y;\theta_0,\gamma_0) = \frac{\partial \log(f(y;\theta_0,\gamma))}{\partial\gamma}\big|_{\gamma=\gamma_0}$ is the part of the score function belonging to the additional parameter γ. Thus, setting $r(y) = U'(y;\theta_0,\gamma_0)\delta$ in relation (A.1) reduces our more general framework in relation (A.3). Moreover, observe that as $n \to \infty$ the true distribution tends to the parametric distribution. Note that because of this fact, the least false parameter θ_0 can actually be treated as the true limiting parameter value in this framework, not just the least false parameter.

We are now going to derive the joint limiting distribution of the parametric and nonparametric estimators of the focus parameter μ. To do that we make almost analogous assumptions as was the case in the chapter 3 derivations.

Assumption A.0.1. *Let $Y_1,\ldots,Y_n$ be iid variables from a distribution with cdf G_n as given in equation* (A.2)*. Let further μ be a one-dimensional focus parameter and θ be the p-dimensional parameter vector of the parametric familiy of distributions with cdf F_θ and limiting true parameter θ_0. Assume the following analogues of assumption 3.1.1:*

$$\mu(\widehat{G}_n) = \mu(G_n) + \overline{\mathrm{IF}}_{\mu,n}(G_n) + o_p\left(\frac{1}{\sqrt{n}}\right), \tag{A.4}$$

$$E_{G_n}[\mathrm{IF}_\mu(Y_i;G_n)] = 0, \qquad E_{G_n}[\mathrm{IF}_\mu(Y_i;G_n)^2] = \nu_n^* \to \nu^* < \infty \text{ as } n \to \infty, \tag{A.5}$$

in addition to

$$\widehat{\theta}_n = \theta_0 + (K^*)^{-1}\overline{U}_n + o_p\left(\frac{1}{\sqrt{n}}\right), \tag{A.6}$$

$$E_{F_{\theta_0}}[U(Y_i;\theta_0)] = 0, \qquad E_{F_{\theta_0}}[||U(Y_i;\theta_0)||^2] < \infty, \tag{A.7}$$

and finally

$$\left.\frac{\partial \mu_F}{\partial\theta}\right|_{\theta_0} \neq 0. \tag{A.8}$$

In addition we assume that

$$\mu(G_n) - \mu(F_{\theta_0}) = \int \mathrm{IF}_\mu(y;F_{\theta_0})\,\mathrm{d}G_n(y) + o_p(||G_n - F_{\theta_0}||_\infty), \tag{A.9}$$

and that the classical Lindeberg conditions of theorem B.2.5 holds for the random variable $(\mathrm{IF}_\mu(Y_i;G_n), U(Y_i;\theta_0)^t)^t/\sqrt{n}$.

Under these assumptions and local misspecified framework, we get a limiting distribution as given in the following lemma:

Lemma A.0.2. *When the relations and conditions of assumption A.0.1 hold, the following limiting distribution appears:*

$$\sqrt{n}\begin{pmatrix}\widehat{\mu}_{\mathrm{np}} - \mu_{\mathrm{true}} \\ \widehat{\mu}_{\mathrm{pm}} - \mu_{\mathrm{true}}\end{pmatrix} \xrightarrow{L} \Lambda^* \overset{d.}{=} N_2\left(\begin{pmatrix}0 \\ \left(\frac{\partial\mu_F}{\partial\theta}\Big|_{\theta_0}\right)^t (K^*)^{-1}q_1 - q_2\end{pmatrix}, \begin{pmatrix}V^*_{\mathrm{np}} & V^*_{\mathrm{pm,np}} \\ V^*_{\mathrm{pm,np}} & V^*_{\mathrm{pm}}\end{pmatrix}\right), \tag{A.10}$$

where

$$\begin{aligned}
\mu_{\text{true}} &= \mu(G_n) \\
q_1 &= \int U(y;\theta_0) r(y)\, \mathrm{d}\upsilon(y), \\
V^*_{\text{pm}} &= \left(\frac{\partial \mu_F}{\partial \theta}\bigg|_{\theta_0} \right)^t (K^*)^{-1} \left(\frac{\partial \mu_F}{\partial \theta}\bigg|_{\theta_0} \right), \\
V^*_{\text{np}} &= \nu^*, \\
V^*_{\text{pm,np}} &= \left(\frac{\partial \mu_F}{\partial \theta}\bigg|_{\theta_0} \right)^t (K^*)^{-1} Q^*.
\end{aligned}$$

Here $(K^*), \nu^*$ *and* Q^* *are limiting analogous of the quantities in chapter 3, i.e.*

$$\begin{aligned}
K^* &= E_{F_{\theta_0}}[U(Y_i;\theta_0)U(Y_i;\theta_0)^t], \\
\nu^* &= \lim_{n\to\infty} \nu^*_n = \lim_{n\to\infty} \mathrm{Var}_{G_n}(\mathrm{IF}_\mu(Y_i;G_n)), \\
Q^* &= \lim_{n\to\infty} Q^*_n = \lim_{n\to\infty} \mathrm{Cov}_{G_n}(U(Y_i;\theta_0), \mathrm{IF}_\mu(Y_i;G_n)).
\end{aligned}$$

Proof. The proof of this limiting distribution can be carried out almost analogous to the proof of lemma 3.1.2. Firstly, we assume that we have shown that the following limiting distribution holds:

$$\sqrt{n} \begin{pmatrix} \widehat{\mu}_{\text{np}} - \mu_{\text{true}} \\ \widehat{\theta}_n - \theta_0 \end{pmatrix} \xrightarrow{L} N_{p+1}\left(\xi^*, \Sigma^*\right), \tag{A.11}$$

where $\xi^* = (0, (K^*)^{-1} q_1)^t$ and Σ^* may be written as a block matrix of the form

$$\Sigma^* = \begin{pmatrix} \Sigma^*_{00} & \Sigma^*_{01} \\ \Sigma^*_{10} & \Sigma^*_{11} \end{pmatrix},$$

where

$$\begin{aligned}
\Sigma^*_{00} &= \nu^*, \\
\Sigma^*_{11} &= (K^*)^{-1}, \\
\Sigma^*_{10} &= (\Sigma^*_{01})^t = (K^*)^{-1} Q^*.
\end{aligned}$$

Then the delta method (theorem B.2.8) may be applied to this limiting distribution with the following transformation function:

$$S_\mu(z,x) = \begin{pmatrix} z \\ \mu_F(x) \end{pmatrix}.$$

This function has Jacobian matrix given by

$$\dot{S}_\mu(z,x) = \begin{pmatrix} 1 & 0 \\ 0 & \frac{\partial \mu_F(x)}{\partial x} \end{pmatrix}.$$

Thus, the delta method gives

$$\sqrt{n}\begin{pmatrix}\widehat{\mu}_{\text{np}}-\mu_{\text{true}}\\ \mu_F(\widehat{\theta}_n)-\mu_F(\theta_0)\end{pmatrix}=\sqrt{n}\begin{pmatrix}\widehat{\mu}_{\text{np}}-\mu_{\text{true}}\\ \widehat{\mu}_{\text{pm}}-\mu_{0,\text{pm}}\end{pmatrix}$$
$$\xrightarrow{L}\Lambda_0^*\overset{d.}{=}N_2\left((\dot{S}_\mu(\mu_{\text{true}},\theta_0))^t\begin{pmatrix}0\\ \xi^*\end{pmatrix},(\dot{S}_\mu(\mu_{\text{true}},\theta_0))^t\Sigma(\dot{S}_\mu(\mu_{\text{true}},\theta_0))\right)$$
$$=N_2\left(\left(\left(\frac{\partial\mu_F}{\partial\theta}\Big|_{\theta_0}\right)^t\begin{matrix}0\\ \end{matrix}(K^*)^{-1}q_1\right),\begin{pmatrix}V^*_{\text{np}} & V^*_{\text{pm,np}}\\ V^*_{\text{pm,np}} & V^*_{\text{pm}}\end{pmatrix}\right),$$

which is almost the limiting distribution we would like to prove. Note now that $\widehat{\mu}_{\text{pm}}-\mu_{\text{true}}=\widehat{\mu}_{\text{pm}}-\mu_{0,\text{pm}}-(\mu_{\text{true}}-\mu_{0,\text{pm}})$. Thus, subtracting the limit of $\sqrt{n}(\mu_{\text{true}}-\mu_{0,\text{pm}})$ from the above distribution should give the distribution we are aiming for. We have from relation A.9 in assumption A.0.1 that

$$\begin{aligned}\sqrt{n}(\mu_{\text{true}}-\mu_{0,\text{pm}})&=\sqrt{n}(\mu(G_n)-\mu(F_{\theta_0}))\\ &=\int\text{IF}_\mu(y;F_{\theta_0})r(y)\,\mathrm{d}\upsilon(y)+o_p\left(1\right)\\ &=q_2+o_p(1)\xrightarrow{P}q_2\quad\text{as } n\to\infty.\end{aligned}$$

since

$$o_p\left(\sqrt{n}\|G_n-F_{\theta_0}\|_\infty\right)=o_p\left(\|R(y)+o_p(1)\|_\infty\right)=o_p\left(1\right),$$

and

$$\begin{aligned}\sqrt{n}\int\text{IF}_\mu(y;F_{\theta_0})\,\mathrm{d}G_n(y)&=\sqrt{n}\int\text{IF}_\mu(y;F_{\theta_0})\,\mathrm{d}(F_{\theta_0}(y)+R(y)/\sqrt{n}+o_p\left(\frac{1}{\sqrt{n}}\right)\\ &=\sqrt{n}\int\text{IF}_\mu(y;F_{\theta_0})\,\mathrm{d}F_{\theta_0}(y)+\int\text{IF}_\mu(y;F_{\theta_0})r(y)\,\mathrm{d}\upsilon(y)+o_p(1)\\ &=\int\text{IF}_\mu(y;F_{\theta_0})r(y)\,\mathrm{d}\upsilon(y).\end{aligned}$$

Thus, by Slutsky's theorem (B.2.6),

$$\sqrt{n}\begin{pmatrix}\widehat{\mu}_{\text{np}}-\mu_{\text{true}}\\ \widehat{\mu}_{\text{pm}}-\mu_{\text{true}}\end{pmatrix}=\sqrt{n}\begin{pmatrix}\widehat{\mu}_{\text{np}}-\mu_{\text{true}}\\ \widehat{\mu}_{\text{pm}}-\mu_{0,\text{pm}}\end{pmatrix}-\sqrt{n}\begin{pmatrix}0\\ \mu_{\text{true}}-\mu_{0,\text{pm}}\end{pmatrix}\xrightarrow{L}\Lambda_0^*+(0,q_2)^t\overset{d.}{=}\Lambda^*,$$

which is the limiting distribution we wanted to prove. What remains now is to validate the assumed limiting distribution in equation (A.11). Using the assumed relations (A.4) and (A.6), we have that

$$\begin{aligned}\sqrt{n}\begin{pmatrix}\mu(\widehat{G}_n)-\mu_{\text{true}}\\ \widehat{\theta}_n-\theta_0\end{pmatrix}&=\sqrt{n}\begin{pmatrix}\frac{1}{n}\sum_{i=1}^n\text{IF}_\mu(Y_i;G_n)+o_p\left(\frac{1}{\sqrt{n}}\right)\\ (K^*)^{-1}\overline{U}_n+o_p\left(\frac{1}{\sqrt{n}}\right)\end{pmatrix}\\ &=\begin{pmatrix}1 & 0\\ 0 & (K^*)^{-1}\end{pmatrix}\sqrt{n}\begin{pmatrix}\frac{1}{n}\sum_{i=1}^n\text{IF}_\mu(Y_i;G_n)\\ \frac{1}{n}\sum_{i=1}^n U(Y_i;\theta_0)\end{pmatrix}+\begin{pmatrix}o_p(1)\\ o_p(1)\end{pmatrix}.\\ &=\begin{pmatrix}1 & 0\\ 0 & (K^*)^{-1}\end{pmatrix}\sum_{i=1}^n X_{ni}+o_p(1),\end{aligned}$$

where $X_{ni} = (X_{ni,1}, X_{ni,2}^t)^t = (\mathrm{IF}_\mu(Y_i; G_n), U(Y_i;\theta_0)^t)^t/\sqrt{n}$. In chapter 3, an analogous situation was solved by applying the usual central limit theorem to the summand. However, since the summand depends on n through data, the standard version of the CLT cannot be applied. Instead we apply the Lindeberg–Feller central limit theorem (B.2.5). By assumption the two triangular Lindeberg conditions stated in theorem B.2.5 holds for $H_n = G_n$ and the defined X_{ni}. Thus, deriving expressions for $E_{G_n}[X_{ni}]$ and $\mathrm{Var}(X_{ni})$ will determine the exact limiting distribution of $\sum_{i=1}^n X_{ni}$ which furthermore gives the limiting distribution in relation (A.11). First we evaluate the expectations

$$
\begin{aligned}
E_{G_n}[X_{ni,1}] &= \frac{1}{\sqrt{n}} E_{G_n}[\mathrm{IF}_\mu(Y_i; G_n)] = 0, \\
E_{G_n}[X_{ni,2}] &= \frac{1}{\sqrt{n}} E_{G_n}[U(Y_i;\theta_0)] = \frac{1}{\sqrt{n}} \int U(y;\theta_0) \left(f(y;\theta_0) + \frac{r(y)}{\sqrt{n}} + o_p\left(\frac{1}{\sqrt{n}}\right)\right) \mathrm{d}\upsilon(y) \\
&= \frac{1}{n} \int U(y;\theta_0) r(y)\, \mathrm{d}\upsilon(y) + o\left(\frac{1}{n}\right) \\
&= \frac{1}{n} q_1 + o\left(\frac{1}{n}\right).
\end{aligned}
$$

Thus $\lim_{n\to\infty} \sum_{i=1}^n E_{G_n}[X_{ni}] = (0, q_1^t)^t$. Furthermore, the variances and covariance are on the form

$$
\begin{aligned}
\mathrm{Var}_{G_n}(X_{ni,1}) &= \frac{1}{n} E_{G_n}[\mathrm{IF}_\mu(Y_i; G_n)^2] = \frac{1}{n} \nu_n^*, \\
\mathrm{Var}_{G_n}(X_{ni,2}) &= \frac{1}{n} \left(E_{G_n}[U(Y_i;\theta_0) U(Y_i;\theta_0)^t] - E_{G_n}[U(Y_i;\theta_0)] E_{G_n}[U(Y_i;\theta_0)^t] \right) \\
&= \frac{1}{n} \int U(y;\theta_0) U(y;\theta_0)^t \left(f(y;\theta_0) + r(y)/\sqrt{n} + o_p(1/\sqrt{n}) \right) \mathrm{d}\upsilon(y) \\
&\quad - \frac{1}{n^2} q_1 q_1^t + o\left(\frac{1}{n^2}\right) \\
&= \frac{1}{n} K^* + \frac{1}{n^{3/2}} \int U(y;\theta_0) U(y;\theta_0)^t r(y)\, \mathrm{d}\upsilon(y) + o\left(\frac{1}{n^{3/2}}\right) \\
&\quad - \frac{1}{n^2} q_1 q_1^t o\left(\frac{1}{n^2}\right) \\
&= \frac{1}{n} K^* + O\left(\frac{1}{n^{3/2}}\right), \\
\mathrm{Cov}_{G_n}(X_{ni,1}, X_{ni,2}) &= \frac{1}{n} E_{G_n}[\mathrm{IF}_\mu(Y_i; G_n) U(Y_i;\theta_0)] = \frac{1}{n} Q_n^*.
\end{aligned}
$$

Thus, the limiting covariance of $\sum_{i=1}^n X_{ni}$ is determined by

$$
\begin{aligned}
\sum_{i=1}^n \begin{pmatrix} \frac{1}{n}\nu_n^* & \frac{1}{n}(Q_n^*)^t \\ \frac{1}{n} Q_n^* & \frac{1}{n} K^* + O\left(\frac{1}{n^{3/2}}\right) \end{pmatrix} &= \begin{pmatrix} \nu_n^* & (Q_n^*)^t \\ Q_n^* & K^* + O\left(\frac{1}{n^{1/2}}\right) \end{pmatrix} \\
&\to \begin{pmatrix} \nu & (Q^*)^t \\ Q^* & K^* \end{pmatrix} = \Sigma'
\end{aligned}
$$

Thus, applying the Lindeberg–Feller central limit theorem (B.2.5) to $\sum_{i=1}^{n} X_{ni}$ gives

$$\sum_{i=1}^{n} \left(X_{ni} - (0, q_1^t)^t \right) \xrightarrow{L} N_{p+1}(0, \Sigma') \Leftrightarrow \sum_{i=1}^{n} X_{ni} \xrightarrow{L} N((0, q_1^t)^t, \Sigma').$$

Now since

$$\sum_{i=1}^{n} X_{ni} = \begin{pmatrix} \frac{1}{n} \sum_{i=1}^{n} \mathrm{IF}_{\mu}(Y_i; G_n) \\ \frac{1}{n} \sum_{i=1}^{n} U(Y_i; \theta_0) \end{pmatrix},$$

we may apply Slutsky's theorem (B.2.6) to obtain the limiting distribution validating relation (A.11). Some simple algebra then verifies that

$$\sqrt{n} \begin{pmatrix} \mu(\widehat{G}_n) - \mu_{\text{true}} \\ \widehat{\theta}_n - \theta_0 \end{pmatrix} \xrightarrow{L} \begin{pmatrix} 1 & 0 \\ 0 & (K^*)^{-1} \end{pmatrix} N_{p+1}((0, q_1^t)^t, \Sigma') = N_{p+1}(\xi^*, \Sigma^*),$$

which is exactly the relation we wanted to validate. □

It should be noted assumption A.0.1 is as very similar to the one stated for the regular situation in chapter 3. The key relations (A.4) and (A.6) shows up from suitable Taylor series expansion under weak regularity assumptions. The perceptive reader may have noticed that relation (A.6) is somewhat different from the corresponding relation for the standard iid situation. Here $K^* = E_{F_{\theta_0}}[U(Y_i; \theta_0) U(Y_i; \theta_0)^t]$ replaces $J = -E_G[I(Y_i; \theta_0)]$.[2] The reason for this change is mainly notational, since for this situation

$$E_{F_{\theta_0}}[U(Y_i; \theta_0) U(Y_i; \theta_0)^t] = -E_{F_{\theta_0}}[I(Y_i; \theta_0)].$$

This follows by some algebra when writing out $I(y; \theta)$ and $U(y; \theta)$ in terms of $f(y; \theta)$ and its derivatives, cancelling terms and interchanging integration and derivation. We omit this proof which can be found in most standard statistical textbooks, like Rice (2007). We end this part of the appendix by giving a corollary with a very helpful limiting distribution

Corollary A.0.3. *When the limiting distribution of equation* (A.10) *in lemma A.0.2 holds, also the following limiting distribution appears:*

$$\sqrt{n}\widehat{b} \xrightarrow{L} N(\xi_b^*, V_b^*),$$

where $\widehat{b} = \widehat{\mu}_{\text{pm}} - \widehat{\mu}_{\text{np}}$, $\xi_b^* = \left(\frac{\partial \mu_F}{\partial \theta} \Big|_{\theta_0} \right)^t (K^*)^{-1} q_1 - q_2$ *and* $V_b^* = V_{\text{pm}}^* + V_{\text{np}}^* - 2V_{\text{pmnp}}^*$.

Proof. The results follows directly from lemma A.0.2 by using that linear transformations of normal distributed variables are also normal distributed (theorem B.2.2). □

[2] The important change here is not what the expectation is calculated with respect to, but the expression inside the expectation.

Appendix B

Useful definitions and results

Here we present a definition and a few theorems which apply in need of in the main thesis. The results are gathered from well-known sources and rewritten to match the notation of this thesis. Proofs of the results may be found in the sources we refer to. Most of the results are standard, but they are nevertheless included here for the sake of completeness. Note also that we have simplified some of the results to better match the situations we need them for. This will be emphasized in the header.

B.1 Definitions

Definition B.1.1. (Functional differential, slightly rewritten from Shao (2003, Definition 5.2))

Let T be a functional on $\mathcal{F}_0$, a collection of cdfs on $\mathbb{R}^p$ and let $\mathcal{D} = \{c(H_1 - H_2) : c \in \mathbb{R}, H_j \in \mathcal{F}_0, j = 1, 2\}$.

(i) A functional T on $\mathcal{F}_0$ is Gâteaux differentiable at $H \in \mathcal{F}_0$ if there is a linear functional L_H on $\mathcal{D}$ such that $\Delta \in \mathcal{D}$ and $H + t\Delta \in \mathcal{F}_0$ imply

$$\lim_{t \to 0} \left[\frac{T(H + t\Delta) - T(H)}{t} - L_H(\Delta) \right] = 0.$$

(ii) Let ρ be a metric on $\mathcal{F}_0$ and suppose that $\|c(H_1 - H_2)\|_* = |c|\rho(H_1, H_2), c \in \mathbb{R}, H_j \in \mathcal{F}_0$ defines a norm on $\mathcal{D}$. A functional T on $\mathcal{F}_0$ is ρ-Hadamard differentiable at $H \in \mathcal{F}_0$ iff and only if there is a linear functional L_H on $\mathcal{D}$ such that for any sequence of numbers $t_j \to 0$ and $\{\Delta, \Delta_j, j = 1, 2, \dots\} \in \mathcal{D}$ satisfying $\|\Delta_j - \Delta\|_* \to 0$ and $H + t_j\Delta_j \in \mathcal{F}_0$, we have

$$\lim_{j \to \infty} \left[\frac{T(H + t_j\Delta) - T(H)}{t_j} - L_H(\Delta_j) \right] = 0.$$

(iii) Let ρ be a metric on $\mathcal{F}_0$. A functional T on $\mathcal{F}_0$ is ρ-Fréchet differentiable at $H \in \mathcal{F}_0$ if and only if there is a functional L_H on $\mathcal{D}$ such that for any sequence $\{H_j\}$ satisfying $H_j \in \mathcal{F}_0$ and $\rho(H_j, H) \to 0$, we have

$$\lim_{j \to \infty} \frac{T(H_j) - T(H) - L_H(H_j - H)}{\rho(H_j, H)} = 0.$$

The functional L_H is called the differential of T at H. Replacing $\rho(H_1, H_2)$ by the general norm $\|H_1 - H_2\|_*$ in (ii) and (iii) give the definitions in terms of the norm.

B.2 Theorems

Theorem B.2.1. (Law of large numbers, part of Ferguson (1996, Theorem 4))

Let $X_1, \ldots, X_n$ be iid with mean ξ and $E[|X_i|] < \infty$. Then the average, $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ satisfies

$$\bar{X}_n \xrightarrow{P} \xi \qquad \textit{(weak version)},$$

and

$$\bar{X}_n \xrightarrow{a.s.} \xi \qquad \textit{(strong version)}.$$

Theorem B.2.2. (Linear transformations of the normal distribution, rewritten from Johnson and Wichern (2007, result 4.2).)

Let $X = (X_1, \ldots, X_q)$ be a r-dimensional random variable, distributed as $N_r(\xi, \Sigma)$. Then, if $a = (a_1, \ldots, a_r)$ is a r-dimensional column vector, then

$$a^t X = a_1 X_1 + \cdots + a_r X_r \sim N(a^t\xi, a^t\Sigma a).$$

Theorem B.2.3. (Convergence in probability implied by convergence in law, Lehmann (1998, Theorem 2.3.4).)

If a sequence of random variables $\{X_n\}$ and sequence of scalars $\{k_n\}$, with $k_n \to \infty$ as $n \to \infty$ satisfies $k_n(X_n - c) \xrightarrow{L} Z$ for some scalar c and some random variable Z. Then

$$X_n \xrightarrow{P} c.$$

Theorem B.2.4. (Multivariate Central Limit theorem, slightly rewritten from Ferguson (1996, theorem 5))

Let $X_1, X_2, \ldots$ be iid random vectors with mean vector η and covariance matrix Σ with all elements finite. If $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, then

$$\sqrt{n}\left(\overline{X}_n - \eta\right) \xrightarrow{L} N\left(0, \Sigma\right).$$

Theorem B.2.5. (Lindeberg–Feller central limit theorem, simplified from van der Vaart (2000, proposition 2.27))

For each n, let $X_{n1}, \ldots, X_{nn}$ be p-dimensional independent random vectors from a distribution with cdf H_n and with finite variance, such that

$$\sum_{i=1}^n E_{H_n}[\|X_{ni}\|^2 \mathbf{1}_{\{\|X_{ni}\|>\epsilon\}}(X_{ni})] \to 0 \qquad \textit{for every } \epsilon > 0,$$

$$\sum_{i=1}^n \mathrm{Cov}_{H_n}(X_{ni}) \to \Sigma.$$

Then

$$\sum_{i=1}^{n} (X_{ni} - E_{H_n}[X_{ni}]) \xrightarrow{L} N_p(0, \Sigma).$$

Theorem B.2.6. (Multivariate Slutsky theorem, slightly rewritten from van der Vaart (2000, lemma 2.8))

Let X_n, X and A_n be random vectors, variables or matrices. If $X_n \xrightarrow{L} X$ and $A_n \xrightarrow{P} A$, where A is constant of the same dimension as A_n, then

(i) $X_n + A_n \xrightarrow{L} X + A$,

(ii) $A_n X_n \xrightarrow{L} AX$,

(iii) $A_n^{-1} X_n \xrightarrow{L} A^{-1} X$ *provided* $c \neq 0$,

provided each of the operations makes sense in terms of the dimensions of X_n, X, A_n and A.

Theorem B.2.7. (Taylor's multivariate theorem, slightly rewritten from Lehmann (1998, theorem 5.5.2))

Let S be a function $\mathbb{R}^r \mapsto \mathbb{R}$ for which the first r partial derivatives exists in a neighborhood of a point a. Then for every point b, we have

$$S(b) = S(a) + \dot{S}(a)(b - a) + o\left(\|b - a\|\right),$$

where $\dot{S}(a) = \left.\frac{\partial S(x)}{\partial x}\right|_{x=a}$.

Theorem B.2.8. (Multivariate Delta method, slightly rewritten from Lehmann (1998, theorem 3.7))

Suppose that

$$\sqrt{n}\left(X_n - \eta\right) \xrightarrow{L} N_r\left(0, \Sigma\right),$$

for some r-dimensional random vector X_n depending on n. If S is a function $\mathbb{R}^r \mapsto \mathbb{R}^s$ which is once differentiable at η and has Jacobian matrix $\dot{S}(\eta)$, then

$$\sqrt{n}\left(S(X_n) - S(\eta)\right) \xrightarrow{L} N_s\left(0, \left(\dot{S}(\eta)\right)^t \Sigma \left(\dot{S}(\eta)\right)\right),$$

provided $\left(\dot{S}(\eta)\right)^t \Sigma \left(\dot{S}(\eta)\right)$ *is positive definite.*

Theorem B.2.9. (Continuous mapping theorem for metric spaces, simplified from van der Vaart (2000, theorem 18.11))

Let $(\mathcal{K}_1, d_1)$ and $(\mathcal{K}_2, d_2)$ be two metric spaces, and $S : \mathcal{K}_1 \mapsto \mathcal{K}_2$ be a function between the spaces continuous almost everywhere for the values of the random variable X of $\mathcal{K}_1$. Let also X_n be a random variable in $\mathcal{K}_1$. We then have that

- *If* $X_n \xrightarrow{L} X$, *then* $S(X_n) \xrightarrow{L} S(X)$.
- *If* $X_n \xrightarrow{P} X$, *then* $S(X_n) \xrightarrow{P} S(X)$.
- *If* $X_n \xrightarrow{a.s.} X$, *then* $S(X_n) \xrightarrow{a.s.} S(X)$.

Theorem B.2.10. (Glivenko–Cantelli, from van der Vaart (2000, Theorem 19.1).)
Let $X_1, \ldots, X_n$ *be iid random variables with cdf* G. *Then* $\|\widehat{G}_n - G\|_\infty \xrightarrow{a.s.} 0$, *where* $\|H_1 - H_2\|_\infty = \sup_x |H_1(x) - H_2(x)|$ *is the supremum norm.*

Theorem B.2.11. (Le Cam's uniform convergence theorem, from Ferguson (1996, Theorem 16(a)))

Let $Y_1, \ldots, Y_n$ *be iid random variable with distribution determined by the cdf* G. *Let* θ *be the one-dimensional parameter vector of some parametric distribution with cdf* F_{θ_0} *and parameter space* Θ. *If*

- Θ *is compact,*
- $D(y;\theta)$ *is a function continuous in* θ *for all* y,
- *There exists a function* $R(y)$ *such that* $E_G[R(Y_i)] < \infty$ *and* $|D(y,\theta)| \leq R(y)$, *for all* y *and* θ,

then

$$Pr\left\{\lim_{n\to\infty} \sup_{\theta\in\Theta} \left| \frac{1}{n}\sum_{i=1}^{n} D(Y_i;\theta) - E_G[D(Y_i;\theta)] \right| = 0 \right\} = 1,$$

i.e. $\sup_{\theta\in\Theta} \left|\frac{1}{n}\sum_{i=1}^{n} D(Y_i;\theta) - E_G[D(Y_i;\theta)]\right|$ *converges almost surely to* 0.

Theorem B.2.12. (Rebolledo's martingale central limit theorem, simplified from Andersen et al. (1993, theorem II.5.1).)
Let $M(t)$ *be a martingale (depending on some sample size* n*) and* $V(t)$ *another process, both of dimension* p. *Let also* $[\cdot]$ *and* $\langle\cdot\rangle$ *denote respectively the optional and predictable variation processes. Assume that*

$$[M](t) \xrightarrow{P} V(t),$$

$$\langle M_{j,\epsilon}\rangle(t) \xrightarrow{P} 0,$$

for all $t \in [0,\tau]$, $j = 1, \ldots, p$, $\epsilon > 0$ *where* $M_{j,\epsilon}$ *is the* j*-th component of the vector of jumps of size larger than* ϵ. *Assuming this holds for* M *working on some suitable filtration, we have that*

$$M(t) \xrightarrow{L} M_G(t),$$

as $n \to \infty$, *where* M_G *is a continuous Gaussian martingale with the property that* $M_G(t) \sim N(0, V(t))$.

Theorem B.2.13. (Lebesgue dominated convergence, rewritten from Schilling (2005, Theorem 11.2))

Let $(B, \mathcal{A}, \mu)$ be some measure space, and S_n a sequence of functions integrable with respect to the measure μ. Let further this function satisfy $\|S_n\|_* \leq \|w\|_*$ for all n and some function w which is integrable with respect to the measure μ and some norm $\|\cdot\|_*$. If $S(x) = \lim_{n\to\infty} S_j$ exists for almost every $x \in B$, then S is integrable with respect to μ and we have

$$\lim_{n\to\infty} \int \|S_n(x) - S(x)\|_* \, \mathrm{d}\mu(x) = 0,$$

and

$$\lim_{n\to\infty} \int S_n(x) \, \mathrm{d}\mu(x) = \int \lim_{n\to\infty} S_n(x) \, \mathrm{d}\mu(x) = \int S(x) \, \mathrm{d}\mu(x).$$

Theorem B.2.14. (Differentiability lemma, extended from Schilling (2005, Theorem 11.5))

Let $(B, \mathcal{A}, \mu)$ be some measure space. Let also A be any open set in $\mathbb{R}^m$ for $m \in \mathbb{N}$. Furthermore, let $S : A \times B \mapsto \mathbb{R}$ be a function satisfying the following three conditions:

- *$S(a, b)$ is integrable with respect to the cdf $H(b)$ for every fixed $a \in A$,*
- *$S(a, b)$ is differentiable with respect to any $a \in A$ for any fixed $b \in B$,*
- *$|\frac{\partial}{\partial b}| \leq w(b)$ for all $(a, b) \in A \times B$ with some positive function $w : \mathbb{R}^m \to \mathbb{R}$ which is integrable with respect to the measure μ.*

Then the function $v(a) : A \to \mathbb{R}^m$ defined by

$$\int_B S(a, b) \, \mathrm{d}\mu(b),$$

is differentiable. In addition differentiation and integration may be interchanged such that

$$\frac{\partial v(a)}{\partial a} = \int_B \frac{\partial}{\partial a} S(a, b) \, \mathrm{d}\mu(b).$$

Theorem B.2.15. (Existence of maximum, from Körner (2003, Theorem 4.44).)

Let K be a compact (closed and bounded) subset of $\mathbb{R}^m$ and $S : K \mapsto \mathbb{R}$ a continuous function. Then we can find k_1 and k_2 in K such that

$$S(k_1) \leq S(k) \leq S(k_2),$$

for all $k \in K$. I.e. both the supremum and infimum of S exists and are attained by values in K.

Bibliography

Aalen, O. (1978), "Nonparametric inference for a family of counting processes," *The Annals of Statistics*, 6, pp. 701–726.

Aalen, O. O., Borgan, Ø., and Gjessing, H. K. (2008), *Survival and Event History Analysis: A Process Point of View*, Springer.

Akaike, H. (1974), "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, 19, 716–723.

Andersen, P. K. and Borgan, Ø. (1985), "Counting process models for life history data: A review (with discussion)," *The Scandinavian Journal of Statistics*, 12, 97–158.

Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (1993), *Statistical Models Based on Counting Processes*, Springer.

Anderssen, S. A., Hansen, B. H., Kolle, E., Steene–Johannessen, J., Børsheim, E., and Holme, I. (2009), "Fysisk aktivitet blant voksne og eldre," Tech. rep., The Norwegian Directory of Health.

Billingsley, P. (1999), *Convergence of Probability Measures*, Wiley, 2nd ed.

Buckland, S. T., Burnham, K. P., and Augustin, N. H. (1997), "Model selection: An integral part of inference," *Biometrics*, 53, 603–618.

Claeskens, G. and Hjort, N. L. (2003), "The focused information criterion," *Journal of the American Statistical Association*, 98, 900–916.

— (2008), *Model Selection and Model Averaging*, Cambridge University Press.

Cramér, H. (1928), "On the composition of elementary errors," *Skandinavisk Aktuarietidskrift*, 11, 141–180.

de Jong, P. and Heller, G. Z. (2008), *Generalized Linear Models for Insurance Data*, Cambridge University Press.

Durbin, J. (1973), "Weak convergence of the sample distribution function when parameters are estimated," *The Annals of Statistics*, 1, pp. 279–290.

Efron, B. and Tibshirani, R. J. (1993), *An Introduction to the Bootstrap*, Chapman & Hall.

Ferguson, T. S. (1996), *A Course in Large Sample Theory*, Chapman and Hall/CRC.

Fernholz, L. T. (1983), *von Mises Calculus for Statistical Functionals*, Springer.

Geisser, S. (1975), "The predictive sample reuse method with applications," *Journal of the American Statistical Association*, 70, pp. 320–328.

Grønneberg, S. and Hjort, N. L. (2008), "The copula information criterion," Statistical research report, Department of Mathematics, University of Oslo.

Hjort, N. L. (1990), "Goodness of fit test in models for life history data based on cumulative hazard rates," *The Annals of Statistics*, 18, 1221–1258.

— (1992), "On inference in parametric survival data models," *International Statistical Review / Revue Internationale de Statistique*, 60, 355–387.

Hjort, N. L. and Claeskens, G. (2003), "Frequentist model average estimators," *Journal of the American Statistical Association*, 98, 879–899.

Huber, P. J. (1967), "The behavior of maximum likelihood estimates under nonstandard conditions," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 221–233.

Hurvich, C. M. and Tsai, C. L. (1989), "Regression and time series model selection in small samples," *Biometrika*, 76, 297–307.

Johnson, R. A. and Wichern, D. W. (2007), *Applied Multivariate Statistical Analysis*, Pearson, Prentice Hall, 6th ed.

Konishi, S. and Kitagawa, G. (1996), "Generalised information criteria in model selection," *Biometrika*, 83, 875–890.

Körner, T. W. (2003), *A Companion to Analysis*, American Mathematical Society.

Lehmann, E. L. (1998), *Elements of Large-Sample Theory*, Springer.

Lien, D. and Shrestha, K. (2005), "Estimating the optimal hedge ratio with focus information criterion," *Journal of Futures Markets*, 25, 1011–1024.

Piessens, R., de Doncker–Kapenga, E., Ûberhuber, C., and Kahaner, D. (1983), *Quadpack: a Subroutine Package for Automatic Integration*, Springer.

Prakasa Rao, B. (1983), *Nonparametric Functional Estimation*, Academic Press.

Rice, J. A. (2007), *Mathematical Statistics and Data Analysis*, Duxbury Press, 3rd ed.

Rohan, N. and Ramanathan, T. V. (2011), "Order selection in ARMA models using the focused information criterion," *Australian & New Zealand Journal of Statistics*, 53, 217–231.

Rubenstein, R. Y. and Kroese, D. P. (2008), *Simulation and the Monte Carlo Method*, Wiley, 2nd ed.

Schilling, R. L. (2005), *Measure, Integrals and Martingales*, Cambridge University Press.

Schwarz, G. (1978), "Estimating the dimension of a model," *The Annals of Statistics*, 6, 461–464.

Shao, J. (2003), *Mathematical Statistics*, Springer, 2nd ed.

Silverman, B. (1986), *Density Estimation for Statistics and Data analysis*, Chapman & Hall.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002), "Bayesian measures of model complexity and fit," *Journal of the Royal Statistical Society, Series B*, 64, 583–639.

Stone, M. (1974), "Cross-validatory choice and assessment of statistical predictions," *Journal of the Royal Statistical Society. Series B (Methodological)*, 36, 111–147.

Sugiura, N. (1978), "Further analysis of the data by Akaike's information criterion and the finite corrections," *Communications in Statistics, Theory and Methods*, A7, 13–26.

Takeuchi, T. (1976), "Distribution of information statistics and a criterion of model fitting," *Suri-Kagaku (Mathematical Sciences)*, 153, 12–18.

Tarima, S. (2011), "Targeted model selection," Statistical research report, Medical College of Wisconsin, submitted to the Journal of Statistical Planning and Inference December 2011.

van der Vaart, A. (2000), *Asymptotic Statistics*, Cambridge University Press.

Varian, H. R. (1974), "A Bayesian approach to real estate assessment," in *Studies in Bayesian Econometrics and Statistics in Honor of L.J. Savage*, eds. Feinberg, S. and Zellner, A., pp. 195–208.

von Mises, R. (1931), *Wahrscheinlichkeitsrechnung*, Franz Deuticke.

Wasserman, L. (2006), *All of Nonparametric Statistics*, Springer.